

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

LEAN AND GREEN

*A breeding strategy to cut fertilizer use in
high-yield cereal crops* **PAGES 563 & 595**

SCIENCE PUBLISHING

PEER REVIEW

*Time to make referee reports
part of the published record?*

PAGE 545

PHYSICS

G WHIZZ

*Precision measurements of
the gravitational constant*


PAGES 562 & 582

STRUCTURAL BIOLOGY

CAUGHT IN THE ACT

*Transcription secrets of
RNA polymerase II*

PAGES 560, 601 & 607

 **NATURE.COM/NATURE**

30 August 2018

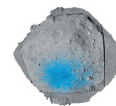
Vol. 560, No. 7720

THIS WEEK

EDITORIALS

WORLD VIEW Reforms would make bullying less likely in science **p.529**

MALARIA Parasites compete to produce drug resistance **p.530**



ASTEROID Japan reveals likely landing sites on rocky Ryugu **p.533**

Opening up peer review

A transparent process to publish referees' reports could benefit science, but not all researchers want their assessments made available.

When *Nature* asks experts to review manuscripts for possible publication, we promise that the reports they send back will be kept confidential. But should we? This week we publish a Comment article (page 545) that comes with a provocative challenge: more journal editors should commit to publishing peer-review reports. Doing so, the authors argue, benefits science. It puts published work in useful context and helps junior scientists to understand how review works.

Nature and the Nature research journals have long welcomed suggestions to make peer review work better for the communities we serve. In 2016, *Nature Communications* started to publish referee reports — with names removed — as long as the authors of the papers agreed.

The reaction has been instructive. For one, it demonstrated that authors in specific fields of the life sciences are more likely to welcome such openness. Take-up from those in other disciplines, including many in the physical sciences, has been much slower. In fact, *Nature Communications* lost several reliable reviewers in chemistry when the referees were told their unsigned reviews would be made public if the author opted for it. They resented not having a say in the process, and felt that their reports would have little value outside the small intended audience.

As such, *Nature* and the Nature research journals have no plans to make publishing referee reports compulsory for all. But we are actively exploring ways to offer it as a wider service in future, when readers, referees and authors say that they want the option.

The desire for transparent peer review is likely to vary across

different communities as they consider the questions involved. Will reviewers shift their focus from a small audience devoted to improving a single manuscript, to persuading a broader audience of their own views of the topic? Will authors be as able to take criticism in their stride, knowing that it will be made public? Will the scientific community get confused by reading criticisms of an earlier draft that no longer apply? Will sections of peer-review reports be presented out of context by campaigners or opponents?

For a publisher, there are other issues to address. An important concern (as when publishing any critical and opinionated material) is the risk of libel in a reviewer's comments or, more commonly, the inclusion in author responses of copyrighted or sensitive third-party material that helps in the assessment of manuscripts but which cannot be made public.

Some clinical-sciences journals now routinely identify reviewers to avoid charges of conflict of interest. By contrast, many journals in the social and political sciences keep authors and reviewers secret during the review process — to encourage frank reports that are not overly awed by prestige or dismissive of under-represented groups. Such double-blind review has been an option on all Nature research journals since 2015.

Nature editors find review reports invaluable. We know that some readers would find them useful as well. We hope that the Comment piece helps to stimulate wider debate. We welcome insights and feedback on this issue from across the scientific spectrum as we continue to align our own practices with the needs of different disciplines. ■

Gravity check

Physicists are stripping uncertainty from the loosest fundamental constant — Big G.

Of the fundamental constants that rule the physical Universe, by far the most perplexing is 'Big G', which quantifies gravitational attraction. The highest and lowest results for G differ by a whopping 0.05%. That might not sound like much, but it's a mad-deniably loose fit for physicists in a world that now routinely measures other constants to ten or more decimal places.

The lack of mastery of G is a mystery. But results reported in *Nature* this week go some way to resolving it (see page 582). The new measurements still don't pin the constant down — in fact, the paper describes two tests from the same laboratory that show a slight statistical difference in their results — but they do offer a way to do so. Because the parallel experiments were performed in the same place, physicists have a chance to narrow down the possible explanations for the discrepancies. (Not only can the set-ups be compared directly, but scientists

already know, for example, that the disparities cannot be down to geographical latitude, or to differences in air density.)

Gravity is the weakest of all known forces (think of how easily a tiny fridge magnet overcomes the downward pull of a planet-sized mass). And getting cash for experiments is tough, because few outside metrology lose sleep over G (most applications rely on relative, rather than absolute, values of gravitational forces). But as they continue to edge towards the truth, Big G researchers can take inspiration from elsewhere in metrology.

The values of some fundamental constants are now so well known that the General Conference on Weights and Measures, which oversees the International System of Units (the SI system), is going to use them in new definitions of the kilogram, ampere, kelvin and mole. The kilogram will no longer be equal to the mass of a physical lump of metal kept in a vault in Paris; instead, it will be defined in terms of Planck's constant, which relates the energy of a wave to its frequency.

Narrowing down these SI constants, such that their uncertainty can now be considered zero, is a triumph of decades-long efforts by labs around the world. It's heartening that such dogged determination continues in the pursuit of Big G. Solving one of the most enduring conundrums of the Universe might not change the world, but it could help us to understand how it works. ■



Research is set up for bullies to thrive

Working conditions in academic labs encourage abusive supervision. It is time to improve monitoring of and penalties for abuse, says **Sherry Moss**.

A young woman contacted me earlier this year to discuss her PhD adviser. He would follow her around the lab, shaming her in others' presence, yelling that she was incompetent and that her experiments were done incorrectly. She wanted nothing more than to minimize contact with him, but she felt trapped. Starting in another lab would mean losing nearly three years of work.

News stories in the past few weeks show this situation to be all too common. Scholars call this kind of workplace bullying abusive supervision. It's a phenomenon I've studied for more than 12 years.

Studies suggest rates of bullying are higher in academic settings than in other workplaces (L. Keashly and J. H. Neuman *Admin. Theory Praxis* 32, 48–70; 2010), but I have no evidence that scientists are more likely than the general population to have characteristics of abusers or their targets. I do think that academic science is a breeding ground for toxic dynamics, mainly because lab heads have so much power over their trainees.

Abusive supervision is more than the occasional lapse into insults, snubs or invasions of privacy. Similar to non-physical domestic abuse, it is defined by sustained hostile behaviour, such as ridiculing, threatening, backbiting and blaming. The 'causes' fall into three categories: characteristics of the target, the supervisor and the situation.

Abusive supervisors often target specific individuals: some pick on their best workers, but poor performers are especially vulnerable. So are those who are different from their adviser, including in gender, ethnicity and sexual orientation. The strongest predictors involve deeper differences, such as working styles, that promote conflict.

Some individuals are more likely to be abusive. Even well-intentioned people in authority are vulnerable to 'power poisoning', which makes them less considerate of others' needs. People who have trouble managing their emotions are more likely to be seen as abusive by employees. So are those with a history of family abuse, or traits such as Machiavellianism (cheating in pursuit of one's interests). And someone who experienced bullying as they rose through the system will often go on to bully.

Stress and perceptions of injustice from above or from external power brokers are also factors. In academic science, lab heads are under pressure from their institutions to publish papers and get grants; that pressure is often passed down to lab members as bullying.

Some supervisors get away with abuse for years. The tendency of universities to take a hands-off approach in the name of academic freedom provides few brakes on outrageous behaviours.

In most workplaces, a bullying boss would see high rates of employee turnover. But in many ways, lab members are captive, making them more vulnerable to abuse. PhD students and postdocs depend on supervisors for publications, funds and letters of recommendation. Changing advisers means years of lost work and, often, damage to a trainee's reputation. The longer a lab member remains, the greater their

commitment to finishing their work under that person, despite abuse.

Abusive supervision has consequences. Those who are abused experience psychological distress, dissatisfaction, emotional exhaustion and depression. It triggers counterproductive behaviours, such as retaliation, aggression towards others and aggression towards the organization — although rarely towards the supervisor. People who are targeted tend to minimize interactions with abusers, although this does not alleviate distress. Social-science experiments suggest feelings of social exclusion, anxiety and stress can lead to unethical choices, such as fudging results (M. Kouchaki and S. D. Desai *J. Appl. Psychol.* 100, 360–375; 2015).

Those experiencing abuse can react in three ways. Most just tough it out, and suffer the psychological consequences. Some change advisers, setting them back in their training but improving their well-being.

After talking to me, the young woman decided to gently confront her adviser. She would tell him that she was uncomfortable with his yelling and would prefer that he speak to her calmly, giving her feedback about what she was doing right and wrong. I never found out whether things improved for her.

Research suggests that only a few confront their bullies, either by speaking up about injustices or explicitly stating how they expect to be treated (B. J. Tepper *et al. Acad. Manage. J.* <http://doi.org/cs82>; 2007). This can improve well-being, but it is risky. Carefully seeking out emeritus faculty members or graduate advisers can help; they might offer insight or be able to intervene with less risk to themselves than some.

And the line between abusive behaviour and

tough, objective and constructive feedback is not always clear.

The best move is never to join a bully's lab. Prospective lab members must ask current ones what it is like to work with the supervisor. Hesitation or responses such as "Being associated with Dr X is an honour, but ..." should give them pause. Too many students look to work with a big name who has lots of publications instead of heeding warnings.

Research institutions must do more to watch for and eliminate abuse. Feedback from lab members should be part of supervisors' appraisals, hiring and promotion. Institutions should conduct exit interviews of lab members, and survey them a few years after leaving. Funders should reward institutions that do this, perhaps with more-favourable indirect costs on grants. In the most egregious cases, institutions should dismiss faculty members or strip abusive supervisors of their right to train PhD students. And the system must create navigable paths for early-career researchers to switch supervisors. When penalties are rare, bad behaviour can thrive. Let's change that. ■

Sherry Moss is a professor of organizational studies at Wake Forest University's School of Business in Winston-Salem, North Carolina, USA. e-mail: moss@wfu.edu

RESEARCH
INSTITUTIONS
MUST
DO MORE
TO WATCH FOR AND
ELIMINATE
ABUSE.

POLICY

Japan gene editing

A government panel in Japan has recommended that the country does not impose regulations on gene-edited organisms. The recommendation, if implemented, would mean that plants or animals that have had their genomes edited at specific locations using tools such as CRISPR would not be subject to the same regulations as genetically modified (GM) organisms. GM organisms, into which foreign DNA has been introduced, cannot be produced without government approval. Although permission would not be required for gene-edited organisms, researchers would have to register them with the government, with the exception of microorganisms that remain in the laboratory. The recommendation is in line with US policy, but contrasts with a July decision by the European Union's highest court that gene-edited organisms should not be exempted from existing regulations. Japan's government intends to convene another panel to review the decision. Currently, no GM crops are grown commercially in Japan.

Power-plant rules

On 21 August, the US Environmental Protection Agency (EPA) revealed its long-promised plan to relax federal limits on greenhouse-gas emissions from power plants. The proposal targets former president Barack Obama's flagship climate regulation, the Clean Power Plan, which sought to reduce carbon dioxide emissions from the power sector to 32% below 2005 levels by 2030. The EPA proposal would allow states to set their own emissions-reduction goals and emphasizes the use of energy-efficiency technologies

at the scale of individual power plants. If finalized and implemented, the proposal will replace the Clean Power Plan, which was introduced in 2015 but blocked by the Supreme Court in 2016 pending a legal review. The EPA will accept public comments on the draft rules for 60 days before finalizing them. See go.nature.com/2mj42pa for more.

'No-deal' Brexit plan

The UK government has released guidance about what will happen to research funded by the European Commission should Britain leave the European Union without making a deal. The document, published on 23 August, warns

that the United Kingdom will be relegated to "third country" status in the European Commission's €80-billion (US\$93-billion) main research-funding programme, Horizon 2020, if no agreement is reached. This will restrict the types of funding researchers in the United Kingdom can bid for. The United Kingdom will leave the EU on 29 March 2019, and scientists could lose access to grants from the European Research Council, which funds basic research, and some parts of Marie Skłodowska-Curie Actions, which facilitate researcher mobility. The guidance reiterates the government's 2016 pledge to underwrite the cost of any

the atmosphere's lowermost 30 kilometres. ESA approved the Aeolus mission in 1999, but development of the satellite's instruments took much longer than expected. The complexity of building a powerful ultraviolet laser that could operate in a vacuum was the main issue. Once Aeolus starts its scientific observations, data from the mission will be incorporated into numerical weather predictions to improve forecasts from national weather agencies.



S. CORVAJA/ESA

Wind-mapping satellite takes flight

The European Space Agency's (ESA's) Aeolus satellite soared into space on 22 August for a three-year mission to monitor wind around the globe. Aeolus is the world's first wind-mapping satellite and launched atop a Vega rocket from Europe's spaceport in Kourou, French Guiana. Mission controllers will spend the next several months calibrating the spacecraft's instruments, including an ultraviolet laser system that will measure the speed and direction of winds in

projects that extend beyond Brexit, if scientists secured the grant before exit day.

RESEARCH

Conductor claims

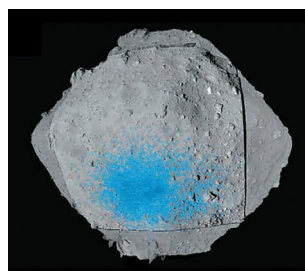
Two teams of physicists have reported hints of superconductivity — the ability of a material to carry a current without resistance — at record-high temperatures. One team, led by Mikhail Erements at the Max Planck Institute for Chemistry in Mainz, Germany, posted a preprint on the arXiv server on 22 August that reported a drop in electrical resistance in a lanthanum-hydrogen compound cooled to below -58°C (A. P. Drozdov

JAVA *et al.* Preprint at <https://arxiv.org/abs/1808.07039>; 2018). If confirmed, it would beat the previous record of -70°C , which the same team obtained in 2015. Then, on 24 August, Russell Hemley of the George Washington University in Washington DC and his colleagues reported preliminary evidence of even more astounding transition temperatures, -13°C and $+7^{\circ}\text{C}$, in a similar material (M. Somayazulu *et al.* Preprint at <https://arxiv.org/abs/1808.07695>; 2018). Both methods required squeezing the samples between diamond tips under extreme pressures. The teams are now working to confirm that the materials are superconductors.

SPACE

Asteroid sites

On 23 August, the Japan Aerospace Exploration Agency revealed the sites on the asteroid Ryugu at which its Hayabusa2 spacecraft will touch down to collect a sample to bring back to Earth — and where it will drop the first two of its four landing probes. Hayabusa2 left Earth in 2014 and reached Ryugu in June. Mission planners faced tough choices for landing sites because the body is strewn with boulders. To minimize



risks for one probe, called MASCOT, mission scientists ran simulations of the surface to produce ten options, and picked one spot (pictured, in blue) on the asteroid's southern hemisphere. The choice reflected various criteria, including average temperatures on the ground. Hayabusa2 will temporarily descend to an altitude of 60 metres to drop MASCOT, and will make its first touchdown in late October, at a site north of Ryugu's equator. The first of three MINERVA-II probes is scheduled to land in late September. See go.nature.com/2wkslms for more.

PEOPLE

Frieden arrest

Tom Frieden, who directed the US Centers for Disease Control and Prevention from 2009 to 2017, was arrested and arraigned in New York City on 24 August. Frieden was charged with forcible touching,

sex abuse and harassment, according to the New York City police department. The charges came after a 55-year-old woman filed a report with police on 7 July accusing Frieden of groping her buttocks in October 2017. On 24 August, a judge ordered Frieden to avoid contact with the alleged victim and released Frieden on his own recognizance. "The allegation does not reflect Dr. Frieden's public or private behavior or his values over a lifetime of service to improve health around the world," a spokesperson for Frieden said in a statement.

Science minister

The Australian government appointed Karen Andrews as the minister for industry, science and technology on 26 August. Andrews was selected by the new prime minister, Scott Morrison, who took up the position two days earlier, following a leadership challenge from another minister that ousted former prime minister Malcolm Turnbull. Turnbull lost the support of his party over an energy policy that sought to reduce the cost of electricity and cut emissions. Andrews, a former engineer, was assistant science minister from September 2015 to July 2016. Science-advocacy

groups say she has been a strong supporter of science and research. Her appointment also puts the position back in the government's inner executive circle, after it was removed in December 2017.

FUNDING

NIH investigation

The US National Institutes of Health (NIH) is investigating about half a dozen institutions with agency-funded researchers who might have failed to disclose monetary support from foreign governments. NIH director Francis Collins spoke about the investigation during a 23 August hearing before a US Senate committee. Earlier that week, the agency sent a letter warning NIH-funded institutions about attempts by foreign governments to influence US research, or to steal intellectual property and other confidential information. The NIH has also set up an advisory panel that will establish methods for accurately reporting all sources of research funding and mitigating security risks to intellectual property. Collins further encouraged institutions to meet with the US Federal Bureau of Investigation about these issues as they pertain to biomedical research.

TREND WATCH

The number of endangered species for which the US government has no recovery plans has grown steadily over the past decade, according to an analysis this month (J. W. Malcom & Y.-W. Li *Conserv. Biol.* <http://doi.org/cs9r>; 2018). The plans, which are mandated under the Endangered Species Act, detail the threats against a species as well as the 'nuts and bolts' strategies that will help the species to recover.

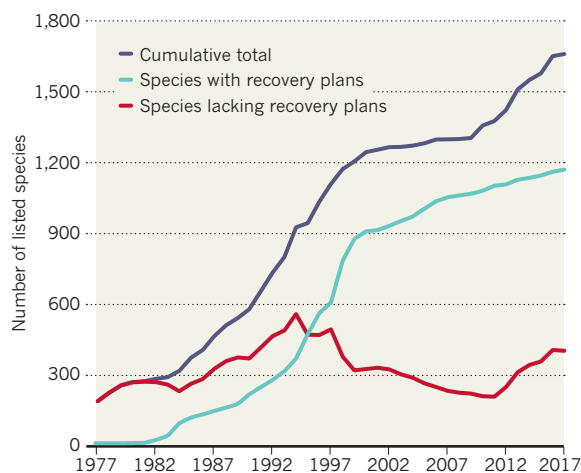
Researchers used publicly available data from the US Fish and Wildlife Service and the National Marine Fisheries Service, the agencies tasked with developing recovery plans, to

work out how many species lacked such safeguards and how that has changed since 1978. They found that of 1,548 eligible species, 24.5% were without a plan.

The analysis also showed that the number of plans has always lagged behind the number of species listed, but that the gap has widened since 2009, when the rate of listings increased. "There isn't enough funding for the services to close that gap," says lead author Jacob Malcom, a conservation biologist at Defenders of Wildlife in Washington DC. And without the plans, he says, people may be unwittingly hindering species' recovery.

ENDANGERED PLANS

Hundreds of plants and animals are protected under the US Endangered Species Act. The drafting of recovery plans has always lagged behind new listings, but the gap has widened since 2009.



NEWS IN FOCUS

PUBLISHING India targets universities in crackdown on predatory journals **p.537**

POLITICS United States woefully unprepared for nuclear strike, say scientists **p.538**

PHYSICS Extraordinary superconductor claim dissected online **p.539**



CONSERVATION How to choose the right trees for forest restoration **p.542**

JESCO DENZEL/VISUM/EYEVINE



One in three deaths worldwide is caused by cardiovascular disease.

FUNDING

Massive grant will go to one heart-research team

British Heart Foundation award is one of the largest single cash pots in medical research.

BY MATTHEW WARREN

A lucky group of researchers will soon walk away with £30 million (US\$39 million) to study the heart and circulatory system — one of the largest single grants for medical research in the world. The British Heart Foundation (BHF) launched the award on 25 August, and it is open to anyone studying heart and circulatory diseases in academia or industry anywhere in the world.

“It’s an absolutely fantastic idea,” says

cardiologist Tim Chico of the University of Sheffield, UK. Cardiovascular disease is responsible for one in three deaths worldwide, and the BHF hopes that providing such a large chunk of money to a single team will accelerate breakthroughs in the fight against the disease. “I think to solve a major problem requires investment at least of this scale,” Chico adds.

Criteria for the award, which is named the Big Beat Challenge, will be published when the application period opens in late 2018.

The grant marks “a very different, radical way of doing things”, says Nilesh Samani, medical director of the charity, which spends more than £100 million a year on cardiovascular-disease research, in grants of up to £3 million. The £30-million award will come on top of the foundation’s usual research investment. Samani says that the aim of the new grant is to fund a big idea that could directly improve the lives of many people.

The BHF has designed its grant to motivate researchers to work across disciplines and ▶

► national boundaries. Cardiovascular problems are often associated with other disorders — for example, kidney and lung diseases — so an approach that cuts across disciplines is important, says cardiac pharmacologist Sian Harding of Imperial College London. “The disease itself doesn’t have boundaries,” she adds. Samani says that applications could also include researchers outside medicine and biology: for example, artificial-intelligence researchers could help to develop tools that predict the risk of cardiovascular disorders.

GLOBAL COLLABORATION

It is also encouraging that the foundation is emphasizing international collaboration and the global burden of cardiovascular diseases, says Amitava Banerjee, a cardiologist and data scientist at University College London. “If we really are talking about global need, then we need to get global data — we can’t be doing studies only in London and Oxford,” he says.

However, Banerjee is concerned that, for all the talk of innovation and transformative research, a grant of this size is likely to go to a team led by well-established, senior scientists who might not be the best source of exciting

ideas. He says that medicine needs to move away from this form of “eminence-based” research, and instead take cues from other industries, in which novel and radical ideas often come from people at a much earlier stage of their career. The BHF says that all applications will need to include diverse teams and no matter who the applicants are, they must be prepared to take a “high-risk, high-reward” approach.

The grant carries with it a substantial amount of money — but it is not without precedent, even in the field of cardiovascular research. In 2015, Google Life Sciences (now Verily) and the American Heart Association announced a \$50-million award for research into preventing coronary heart disease. Pharmaceutical company AstraZeneca later climbed on board as well, adding an extra \$25 million to the pot.

The winner of that grant was Calum MacRae, a cardiologist at Brigham and Women’s Hospital in Boston, Massachusetts,

who is looking for early markers of coronary heart disease. Compared with smaller pots of funding, large grants can force researchers to think about problems in completely different ways, MacRae says. The grant has enabled his team to work at a faster pace and more collaboratively than might have been possible with more traditional forms of support, he says. “Diversity in funding is as important as diversity in ideas.”

Another UK-based charity, Cancer Research UK (CRUK), also awards “Grand Challenge” grants of up to £20 million to address specific problems. Last year, four teams received funding for projects such as creating virtual-reality maps of tumours and finding ways of preventing unnecessary breast-cancer treatment. Ten teams have been shortlisted for a second round of funding.

The BHF consulted MacRae and CRUK when planning their new award, says Samani. But they decided not to restrict the scope, and instead give researchers leeway to pitch any project related to heart and circulatory disease. “We really trust the research community to come up with the best ideas,” he says. “I’m not aware of any other major grant of this scale which is that open.” ■

“If we really are talking about global need, then we need to get global data.”

POLITICS

Trump science-adviser pick hedges on climate change

Meteorologist Kelvin Droegemeier offered few clues to his views on the topic to US lawmakers.

BY SARA REARDON

Kelvin Droegemeier — President Donald Trump’s nominee for science adviser — revealed little about his stance on climate change during his nomination hearing before a US Senate committee on 23 August. Some experts attribute his elusiveness to deftly manoeuvring a politically sensitive topic, rather than doubts about the science.

The researcher, a meteorologist whom Trump nominated to lead the White House Office of Science and Technology Policy (OSTP) on 31 July, told committee members that science should be conducted without political interference or influence. “I am absolutely firm on the point,” Droegemeier said.

But he equivocated on whether views that are in the minority, such as doubts about the human role in climate change, should be included in policymaking decisions. “Science never provides immutable evidence about anything,” he said. “I think science is the loser



Climate change is likely to exacerbate extreme weather.

STAN GROSSFELD/THE BOSTON GLOBE/GETTY

when we tend to vilify and marginalize other voices. We need to have everyone at the table.”

When pressed by Republican and Democratic committee members about climate change, Droegemeier offered little, other than saying that bringing the weather and climate-modelling communities together could improve forecasts.

If confirmed, the meteorologist would join an administration that has sought to cut climate-change programmes at the Environmental Protection Agency and roll back federal regulations on greenhouse-gas emissions.

Scientists were largely encouraged when Trump nominated Droegemeier to lead the OSTP, which helps to coordinate science policy and spending between federal agencies. And Neal Lane, a physicist who served as science adviser to former president Bill Clinton, remained optimistic. “No one in Congress is going to say extreme weather events are not important,” he said. And linking those episodes

with climate science is vital, Lane added. “There’s nobody better to do that than Kelvin Droegemeier.”

Other questions from lawmakers focused on scientific competition from China and on sexual harassment in research.

“We need to make sure we are the strongest research centre in the world,” said Droegemeier. And although welcoming foreign researchers is an important part of science in the United States, he said, it should be done with care.

He also spoke in favour of a recent National Science Foundation (NSF) policy that requires institutions to report agency-funded researchers who are found to have committed sexual harassment. “We owe all scientists a safe place to work,” Droegemeier said. If confirmed to lead the OSTP, he said, he will turn the attention of all agencies under his purview to this issue. He also plans to focus on increasing representation of women and people from under-represented groups in science.

“I think it’s a bright day for science,” said Lane, who had written to the Senate committee in support of Droegemeier’s nomination.

In his opening remarks, Republican Senator James Inhofe of Oklahoma said that Droegemeier has impressive scientific qualifications. The meteorologist was vice-president for research at the University of Oklahoma in Norman from 2009 to 2018. He stepped down from his position on 20 August, in advance of his confirmation hearing.

Droegemeier also served on the National Science Board, which oversees the NSF, under former presidents Barack Obama and George W. Bush. He is the current secretary of science and technology for Oklahoma.

The Senate committee will vote on 29 August on whether to advance Droegemeier’s nomination to the full Senate. If a majority votes for his confirmation, he will be the first non-physicist to take the reins at the OSTP since Congress established the office in 1976. ■

PUBLISHING

India targets fake journals

The government tells universities to stop promoting predatory publications.

BY SUBHRA PRIYADARSHINI

Most academics regard predatory journals as an irritant — if not a threat — to science. But in India, some universities have recommended the inclusion of such publications in the country’s ‘white list’ of approved journals. Now the government is cracking down on this practice, which scientists say came about as a result of perverse government incentives.

“We will end this menace of predatory journals,” Prakash Javadekar, the minister responsible for higher education, told parliament last month. Universities now have until the end of August to revise their recommendations for the journal white list to avoid predatory publications, which actively solicit manuscripts and charge authors hefty fees without providing the services they advertise, such as editing and peer review.

Predatory journals are a problem because research funding is wasted on deceptive publishers that don’t deliver what they promised. A major international journalistic investigation, published last month in multiple media outlets, estimated that the number of papers put out by five major predatory publishers has tripled since 2013 — to about 175,000 articles.

Many publishers that host suspected predatory journals are based in India. And multiple studies have found that a high proportion of articles in such journals come



India’s universities minister Prakash Javadekar has promised to end the “menace” of predatory journals.

from academics in the country^{1,2}.

Many Indian academics blame this situation on the nation’s system for assessing academic performance. In 2010, India’s higher-education regulatory and funding agency, the University Grants Commission (UGC), introduced a system for evaluating academics called the Academic Performance Indicator, which places considerable weight on the number of research publications. Universities must

use the indicator to hire and promote faculty members. But scientists have complained that this encourages academics and universities to focus on the quantity of publications, rather than their quality.

To reduce the practice of publishing in sub-standard journals, the UGC released a white list of approved journals in January 2017. The list contained approximately 32,000 publications indexed on science-citation ►

► databases such as Web of Science and Scopus, as well as more than 5,000 publications recommended by universities. But researchers quickly pointed out that it also included predatory journals.

Virander Singh Chauhan, who chairs the UGC committee that assesses and accredits higher-education institutions and who oversaw the list, says that the predatory journals had been recommended by some universities, and that the UGC had learnt of this only later. Unless universities stop doing this, “nothing can get rid of fake journals in India”, says Chauhan. Currently, he says, universities can simply recommend journals, and make minimal effort to check a publication’s quality.

In May, the UGC removed 4,305 journals from the list on the grounds of poor quality, or because incorrect or insufficient information about the journal had been provided. (The group will update the list with universities’ revised recommendations.) Chauhan says that introducing stricter criteria for registering journals on the UGC list would reduce the number of predatory publications.

Ajit Kembhavi, an astrophysicist at the Inter-University Centre for Astronomy and Astrophysics in Pune, says the government’s plan to crack down on university-proposed journals is a good first step, but that the bigger problem is how universities are evaluated and funded.

A more permanent solution would be to decouple academic assessments from a researcher’s number of publications, says Kembhavi. He adds that more also needs to be done to promote greater awareness of predatory journals among academics in India and to educate them about research ethics.

In China, where some universities reward academics on the basis of the number of publications, the government is working on a blacklist of journals it deems to be of poor quality, or set up only for profit. Research published in these journals will not count towards promotion or grant applications, and the authors will also receive a warning.

Bhushan Patwardhan, a biologist at Savitribai Phule Pune University and a vocal critic of dubious publishing practices, says the Indian government should also show zero tolerance towards academics who publish in these journals. There are currently no repercussions for those who do this. He says the government should introduce rules similar to regulations introduced to detect and punish plagiarism at universities, which came into effect in July. “If faculty members are allowed to get away with such practices, what would stop them from doing this again?” says Patwardhan. ■

1. Xia, J. *et al.* *J. Assoc. Inform. Sci. Technol.* **66**, 1406–1417 (2015).
2. Shamseer, L. *et al.* *BMC Med.* **15**, 28 (2017).



BETTMANN/GETTY

A nuclear blast can cause mass death and damage across a wide area.

PUBLIC HEALTH

US unprepared for nuclear attack

Growing threat from North Korea rattles scientists who study disasters and public health.

BY SARA REARDON

The United States is not prepared to deal with the aftermath of a major nuclear attack, despite North Korea’s efforts to develop nuclear weapons and the increasing tensions between nations overall. That was the blunt assessment of public-health experts who participated in a meeting last week on nuclear preparedness, organized by the National Academies of Sciences, Engineering, and Medicine.

The gathering is “an acknowledgement that the threat picture has changed, and that the risk of this happening has gone up”, says Tener Veenema, who studies disaster nursing at Johns Hopkins University in Baltimore, Maryland, and who co-chaired the conference in Washington DC.

Since the fall of the Soviet Union in 1991, the United States’s research and preparedness efforts for a nuclear strike have focused largely on the possibility of a terrorist

attack with a relatively small, improvised 1-kilotonne weapon or a ‘dirty bomb’ that sprays radioactive material.

But North Korea is thought to possess advanced thermonuclear weapons — each more than 180 kilotonnes in size — that would cause many more casualties than would a dirty bomb (see ‘Damage estimate’).

“We’re back to people saying, ‘We can’t deal with this.’”

“Now that thermonuclear is back on the table, we’re back to people saying, ‘We can’t deal with this,’” says Cham Dallas, a

public-health researcher at the University of Georgia in Athens.

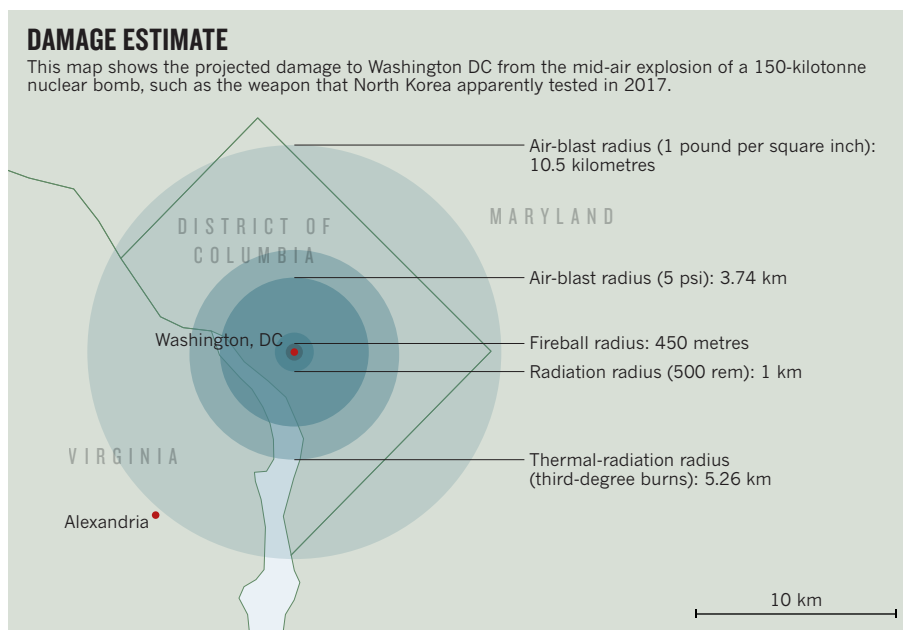
Veenema says that the science academies decided to do a study in November 2017, three months after North Korean leader Kim Jong-un threatened to launch a nuclear weapon at the US territory of Guam. The academies wanted to bring together the different government, academic and private sectors that would be

involved in the medical response to a nuclear attack. The academies' committee plans to release a report in December that lays out how the United States could plug the gaps in its response capabilities.

The US government's spending on nuclear-weapons research and response has dropped drastically over the past few decades — as has the number of health workers with training in radiation medicine and management. According to a 2017 study by Dallas, more than half of emergency medical workers in the United States and Japan have no training in treating radiation victims (C. E. Dallas *et al. Front. Public Health* 5, 202; 2017).

The same study suggests that even trained medical professionals might be too frightened to enter a nuclear-fallout zone or to treat radiation victims at the scene — Dallas's group found that 33% of medical professionals said they would not be willing to respond in such a scenario.

Compounding these concerns, treatments for radiation exposure and burns might not be available in sufficient quantities in the aftermath of a nuclear attack. James Jeng, a burns surgeon at Mount Sinai Health System in New York City, says that the detonation of a nuclear bomb can leave behind hundreds of thousands of burn victims. The best treatment for such injuries is skin grafting, he says, but



SOURCE: NUCLEARSECURITY.COM/NUKEMAP

there are only about 300 burn surgeons in the United States who know how to perform the procedure. It might also be difficult to quickly transport enough donor skin to treatment sites, Jeng adds.

North Korea's threat to Guam last year made clear to public-health officials there

how limited their response capabilities are, says Patrick Lujan, emergency-preparedness manager for the Guam Department of Public Health and Social Services. Guam, an island of 163,000 people, has only three hospitals and no burns units. "We realized there's just so much you can do, being on an island," Lujan says. ■

PHYSICS

Social-media storm dissects superconductivity claim

Thrill over potential high-temperature superconductor reached fever pitch, then died away.

BY DAVIDE CASTELVECCHI

It was an explosive claim: the discovery of a superconducting material that can carry electricity with almost no resistance in normal conditions. The purported finding — announced by two physicists¹ last month — sparked a rush of replication efforts. But independent researchers have grown sceptical as they have dissected the claim, in a process that played out mostly on social media.

"All these researchers who normally do not discuss on a single platform have come together and discussed this," says Pratap Raychaudhuri, who studies low-temperature physics at the Tata Institute of Fundamental Research in Mumbai, India. He led a discussion of the results on Facebook. "I think the self-correcting mechanism of science — the ruthless scrutiny of the community — has

worked extremely well," he says.

Most superconducting materials identified so far work only at much lower temperatures, often close to absolute zero. The highest seen yet is -70°C , reported² in 2015 — and that compound is superconducting only at extremely high pressures. (Just last week, the same laboratory posted³ a preprint on the arXiv server describing a new record, -58°C , for superconductivity at high pressure, but that result has not yet been confirmed.) In a preprint posted¹ on 23 July, Dev Kumar Thapa and Anshu Pandey of the Indian Institute of Science in Bangalore (IISc) described a material made from gold and silver that became superconducting at a balmy -37°C , and at normal ambient pressure.

"It was a remarkable claim, so there was lots of interest," says Raychaudhuri. Several laboratories quickly leapt into action to try to

replicate the results. But their efforts were frustrated, because the preprint did not provide the details needed to manufacture the gold-silver material, and because Thapa and Pandey declined requests to share their samples.

Thapa and Pandey told *Nature's* news team that they would not comment on their research while their paper is under review at a journal. Pandey said that they are having their results validated by independent experts, and that they will announce the results of the validation in the appropriate forum as soon as possible.

TWITTER CHATTER

Brian Skinner, a theoretical physicist at the Massachusetts Institute of Technology (MIT) in Cambridge, began studying the preprint soon after it came out — and eventually chronicled his findings in a widely shared Twitter thread. Although superconductors are not ►



Superconductors can levitate objects — but require very low temperatures.

his speciality, the excitement surrounding the paper piqued his curiosity. He noticed that one of the preprint's figures contained curves of data points that were surprisingly free of random background noise at relatively warm temperatures, but became noisier below the temperature at which the material transitioned to a superconducting state. "Usually, they look smooth on both sides, or dirty on both," Skinner says.

When he zoomed into the picture, Skinner was even more surprised: the graphic included

several data sets in which the experiment was run in slightly different conditions, and the patterns of noise seemed very similar for each run. But noise is, by nature, random. He went to discuss his observation with an expert on superconductors at MIT, who agreed that the pattern was odd. And in the following days, Skinner had conversations with many other researchers.

Repeated patterns of noise alone do not necessarily mean that the data are faulty or intentionally fabricated, Skinner says, but he still wanted the broader community to know

about his concerns. So on 8 August, Skinner submitted a two-page response⁴ to the preprint on arXiv. The post, which he mentioned on Twitter, prompted a viral response, with more than 3,600 shares and countless online mentions.

A separate MIT group — led by experimentalist Mingda Li — that had also been attempting to replicate the results took note of Skinner's post, and Li became concerned. "Fluctuations really shouldn't be that identical," he says. His group decided to call off their replication attempts.

On Facebook, Raychaudhuri gave a possible explanation for the repeating data patterns, and said that to get to the bottom of the story, the authors need to share their data. But although Raychaudhuri is not convinced by the claims, the affair has provided an opportunity to show science in action.

As for the claims, says Li, "if the authors don't provide any new experimental measurements, this will gradually go away". ■

1. Thapa, D. K. & Pandey, A. Preprint at <https://arxiv.org/abs/1807.08572> (2018).
2. Drozdov, A. P., Erements, M. I., Troyan, I. A., Ksenofontov, V. & Shylin, S. I. *Nature* **525**, 73–76 (2015).
3. Drozdov, A. P. et al. Preprint at <https://arxiv.org/abs/1808.07039> (2018).
4. Skinner, B. Preprint at <https://arxiv.org/abs/1808.02929> (2018).



HOW TO REBUILD A FOREST

As projects to restore woodlands accelerate, researchers are looking for ways to avoid repeating past failures.

BY RACHEL CERNANSKY

When the Philippines opened its first school of forestry in 1910, the institute's leaders hatched a plan to restore degraded woodlands surrounding the campus outside Manila. They planted dozens of tree varieties, both native and exotic. In 1913, the school received 1,012 mahogany (*Swietenia macrophylla*) seeds from a botanical garden in Calcutta, India, and started growing them around the grounds. The American hardwood became such a staple of reforestation efforts in the country that it spread throughout natural areas, so much so that it eventually proved a nuisance. The trees create veritable green deserts:

their tannin-rich leaves are unpalatable to local animals and seem to stifle the growth of other plants where they fall. They also produce seeds annually, giving them an advantage over native hardwoods, which do so at intervals of five years or more.

It's hardly history's only forestry folly. "The whole notion of what species should be used in restoration tends not to receive, I would say, adequate attention," says Douglas McGuire, coordinator of the Forest and Landscape Restoration Mechanism at the Food and Agriculture Organization of the United Nations in Rome.

Many projects fail because they choose the wrong trees, use too few species or are not managed for the long term. Foresters and ecologists are realizing that for restoration efforts to succeed, they need to think more broadly — about matching trees to their location, about the effects on nearby insects and other animals and about relationships with soil and the changing climate. In other words: the ecosystem.

Scientists are now testing and comparing strategies that range from letting nature take its course, to forest-management approaches that look a lot like farming. There is no one-size-fits-all solution, but the work exposes some philosophical friction. Ecologists seeking to increase biodiversity might champion a broad range of species, whereas sustainable-development advocates could back exotic fruit-bearing trees that benefit local people. And researchers seeking to mitigate climate change might push for a single fast-growing variety.

"There've been different attitudes about what the goal of restoration is," says Robin Chazdon, a forest ecologist at the University of Connecticut in Storrs. "There is also some attempt to reconcile, which is very promising."

There is room for growth — a lot of it, in fact. A 2011 analysis suggested that some 2 billion hectares of land, an area larger than South America, is suitable for restoration (see 'Green expectations'). Much of this land has been deforested or degraded as a result of human activity. And many countries and organizations have made promises in the past decade to help fill that area. There are pledges to plant billions or even trillions of

A Brazilian nursery grows seedlings to support reforestation efforts.

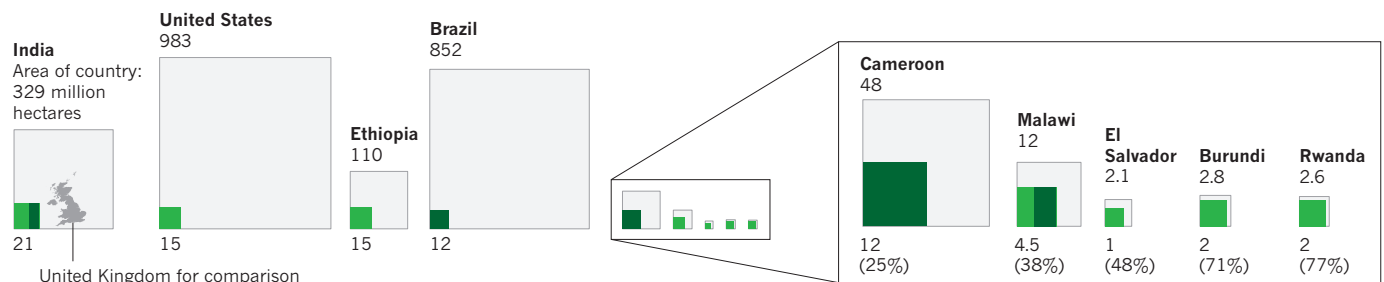
MINT IMAGES/AIURA

GREEN EXPECTATIONS

Roughly 2 billion hectares of land could be suitable for forest landscape restoration, according to a global analysis of current status and human pressures. Although it doesn't offer local prescriptions, the exercise broadly outlines areas of land appropriate for wide-scale restoration, remote areas not amenable to direct management and land that could support a mosaic of tree cover and small-scale farming. It excludes urban areas, intensive agriculture and already forested lands.

AMBITIOUS GOALS

Forty seven countries have pledged to restore degraded lands as part of the Bonn Challenge for 2020 and 2030. Here are some of the largest commitments by land area (left) and by proportion to the total size of the country (right).



trees, and regional programmes such as Africa's Great Green Wall, which would surround the Sahara Desert with vegetation. China has set some of the most ambitious national targets. It is aiming to plant 6.7 million hectares' worth of trees — roughly the size of Ireland — this year alone.

But some key deadlines are looming. The Bonn Challenge, established in 2011, for example, aims to restore 150 million hectares by 2020, and another 200 million in the subsequent decade. It has received ample commitment from countries around the world, but the strategies aren't always backed by evidence, and measures of success are still being defined. As conservation efforts move forwards, scientists say, it's imperative to look at the leading strategies. "There's a big risk in this restoration movement of big promises, big targets and a time frame that's really tight," McGuire says.

LET NATURE TAKE ITS COURSE

When people think of reforestation, they often think of planting trees. But some ecologists argue that the best way to repopulate a forest is to leave it alone. In the 1980s, Daniel Janzen and his partner Winnie Hallwachs, both biologists at the University of Pennsylvania in Philadelphia, developed a plan to reforest a small national park in Costa Rica that had been carved out of a former ranch. It was covered in African grasses that were intentionally burned during the dry season. The pair, along with partners including the government, employed local people to stop the fires and help guard the land. Over time, what had resembled overgrown African savannah became a tropical forest with rain trees (*Samanea saman*), guanacaste (*Enterolobium cyclocarpum*), hog plums (*Spondias mombin*) and other native trees. And with the help of donors and local workers, it grew.

Today, the Guanacaste Conservation Area, a World Heritage Site with more than 100,000 hectares of land, is seen as one of the best examples of this approach to restoration, known as natural regeneration. Janzen is a vocal proponent of the strategy. Take away the assault, and "nature takes care of the restoration," he says. "Organisms like to get their land back."

But natural regeneration won't work everywhere. There are countless areas around the world that are much more degraded than Guanacaste. In

some places, soil nutrients are depleted, and there are no seeds or seedlings from native species to populate the space. Even with the political will to protect such regions, forests are unlikely to regrow.

That is where more aggressive efforts are needed, and conservationists are exploring different strategies. In Thailand, Stephen Elliott, research director for Chiang Mai University's Forest Restoration Research Unit, has been restoring local forest with native species for decades. He's followed a framework-species approach, which involves planting enough species to start attracting pollinators and seed dispersers. The key, he says, is getting the canopy to close quickly enough — by the second or third year — to prevent weeds from taking over.

Nigel Tucker, who helped to establish the framework-species approach in Australia in the 1990s, says that he noticed early on that some plants had an outsized role in supporting a thriving ecosystem. Take fig trees (*Ficus* spp.): in tropical forests around the world, they produce regular fruit crops that birds, bats and primates rely on — particularly during dry periods — and their foliage is an important food source for other animals. All of that helps with pollination and seed dispersal, which encourages regeneration of the forest. "In my work locally, figs always comprise 10% of any planting, and we plant as many fig species as possible," Tucker says.

Another strategy, known as applied nucleation, involves planting small clusters, or 'nuclei', of trees throughout a clearing. The goal is for these to gradually close in on each other, as the nuclei attract seed dispersers. Karen Holl, a restoration ecologist at the University of California, Santa Cruz, has studied this approach in Costa Rica and elsewhere. It can be just as effective as planting a whole area with trees, she says, but it requires fewer resources, and the outcome is a more varied-looking landscape.

Chazdon has been working with colleagues to write a review that compares how the different approaches affect timber production, wildlife populations, water and sediment retention, and other factors. But she is struggling to do so because, she says, there aren't many studies to review. "We don't have a lot of evidence. We have perceptions," she says. "The basis for decision-making is not very scientific at this point."

COOPERATIVE APPROACHES

Despite forestry blunders such as the Philippines' mahogany problem, researchers still debate whether restoration efforts must rely entirely or predominantly on native species. A growing number of efforts are showing that integrating exotic commercial species with native ones can produce promising results both for ecosystems and for economies. Species such as eucalyptus (*Eucalyptus globulus*) and pine (*Pinus* spp.) can grow quickly, and in very degraded soils; most of the native species that are being lost in forests around the world do neither. Planting them together means that the faster-growing trees — chosen because they can't spread on their own — can provide a canopy for the slower ones, giving them a helping hand. The canopy species can also be a source of income for communities or a way to appeal to timber companies to participate in restoration projects that promote species diversity. Restoration ecologist Pedro Brancalion at the University of São Paulo's Tropical Forestry Lab in Brazil is collaborating with a wood-pulp company to plant eucalyptus trees alongside native species in the Atlantic Forest and later harvest the eucalyptus. The approach has generated enough revenue to offset most of the project's costs.

Native species can benefit economies, too. Another effort Brancalion is involved with leans heavily on juçara (*Euterpe edulis*), a threatened relative of the better-known açai that also produces an edible fruit. Juçara trees are planted wherever people see fit: in home gardens, along the small dirt roads that connect villages, in fragments of remaining forest and in agroforests — where trees or shrubs are integrated with other food crops or with pastureland. A project known as the Juçara Network has also revived cultural appreciation for the fruit, which is now the focus of a national gastronomic festival and a key source of income for many small farmers.

Chazdon and others say that in heavily populated areas, agroforestry seems like a good idea because it can provide food. "That will be a strong motivating factor for people to become involved and to make the restoration successful," she says.

It has been catching on in parts of Africa. Alex Munyao, a farmer in eastern Kenya, learned how to care for seedlings and graft trees at a training programme in 2013 hosted by the Nairobi-based World Agroforestry Centre, or ICRAF. He convinced the ICRAF team to establish a nursery that grows avocados (*Persea americana*) originally from Mesoamerica, kei apples (*Dovyalis caffra*), which are native to southern Africa, and a handful of other fruits. He has now sold more than 30,000 seedlings to other farmers and to local government officials for restoration projects. He has also donated some to local schools, and helps people in the community to graft their own local avocado trees with improved varieties.

Stepha McMullin, who runs the Fruiting Africa programme at ICRAF in Kenya, says that because people like Munyao are spreading the word, such training has been able to reach 10,000 or more farmers. The programme has distributed enough seedlings to plant trees on more than 500 hectares of farmland. It does include exotic species, partly because fruits such as mangoes and papayas often have higher market values, but farmers are learning the value of some native varieties, too.

The desert date (*Balanites aegyptiaca*), for example, was once common in the wild in much of Africa's dry lands and its fruit was nutritious and popular with children, but many farmers had cleared these trees from their land to make way for other crops. When McMullin's team approached farmers about planting — or simply sparing — desert dates, "they were very surprised and even laughed at the thought", she says. But after learning about the health benefits, particularly for children, more families have opted to preserve and plant the trees.

A QUESTION OF ORIGIN

In an effort to support restoration programmes elsewhere and on a larger scale, McMullin's colleagues are developing supplies of seeds and seedlings, maintaining gene banks and sequencing the genomes of indigenous trees and other crops. Their work deals with one of the

problems that could block major restoration efforts in different parts of the world.

"Where's the planting material going to come from? That's one big bottleneck," says Ramni Jannadass, a genetic-resources specialist who oversees ICRAF's Tree Diversity, Domestication and Delivery project.

In May, Bioversity International and other organizations released a report analysing the seed-supply systems in seven Latin American countries, focusing on the government and research agencies involved in restoration (see go.nature.com/2p3gmke); none paid much attention to the genetic origins of the seeds or the diversity of the native species available.

Brazil is an exception to that trend, having established thriving nurseries for native seedlings. It also has laws requiring landowners in the Amazon to maintain native vegetation on a certain amount of their property — although these laws have had mixed success. They were not enforced for a long time, and by some estimates, deforestation has increased over time, not declined.

Asia is arguably the region most neglected by global efforts to increase diversity in restoration and to study native species. Christopher Kettle,

Bioversity International's director for forest genetic resources and restoration in Rome, says that the need for infrastructure — things such as mechanisms for collecting and storing seeds, and nurseries to raise seedlings — might be most desperate here because many trees are 'masting' species, which don't produce seeds every year. People need to be ready. "Otherwise, you miss the boat, you lose all the seed and you've got to wait another seven years," says Kettle. "This is a really, really critical issue for restoration in Southeast Asia, because many of the

most important timber species and tree species — the ones that will lock up the most carbon — they're all masting species."

Climate change is a driving factor in the push to restore forests, but it also raises questions, such as where trees can thrive in the future. John Stanturf, a forest ecologist and research-group coordinator at the International Union of Forest Research Organizations in New York, sees promise in the concept of assisted migration, or moving plants to where they can survive today and thrive in the future. He and his colleagues last year collected seeds from Iran's Caspian forests, and brought them to Denmark. The Iranian trees are adapted to heat and droughts, but also related to the Danish species. Stanturf plans to test whether the introduction increases genetic diversity, resistance and resilience in the native trees.

Climate change is also expected to alter relationships between trees, insects, diseases and other forest species. "Insects that today are a minor problem may become a major problem if they can produce three or four generations in a year," says Stanturf. This remains a significant knowledge gap. "We know enough to know that this is a concern, but we don't know enough about how to respond to it yet. That's a great area to be doing research." So is soil, says Cindy Prescott, a forest ecologist at the University of British Columbia in Vancouver. "If you don't look at the soil at the start, you can spend a lot of money and time putting in species that aren't going to survive there."

With so much research left to do, leaders in the field have been doing some soul-searching, and acknowledging that restoration can be motivated by — and designed to meet — different needs. "When you talk about conservation or restoration, the first question has to be restoration by whom, for whom?" says Janzen.

The question can have more than one answer. Much of the global funding for restoration is dedicated to developing it as a tool to mitigate climate change, notes Brancalion. "But if you ask a farmer in Brazil if he or she is concerned about climate change, they would say, 'No, I am concerned about water,'" he says. Their interests as stewards of the land need to be better integrated with those who have the money to support restoration.

That has been the strongest lesson of all for Chazdon. Restoration is about more than what gets planted in the ground, she says. "Yes, it's about forests, but it's really about people. They are the agents of restoration." ■

Rachel Cernansky is a science journalist based in Denver, Colorado.

COMMENT

HISTORY Seventy-five years since physicist Schrödinger made waves in biology **p.548**



TECHNOLOGY Why are governments in favour of digital insecurity? **p.550**

PUBLISHING Preprints — what is good for science is good for the public **p.553**

OBITUARY Burton Richter, charm-quark Nobel laureate, remembered **p.554**

ILLUSTRATION BY DAVID PARKINS



Publish peer reviews

Jessica K. Polka and colleagues call on journals to sign a pledge to make reviewers' anonymous comments part of the official scientific record.

Long shrouded in secrecy, the contents of peer review are coming into the open. In the past decade, outlets such as *eLife*, *F1000Research*, *Royal Society Open Science*, *Annals of Anatomy*, *Nature Communications*, *PeerJ* and EMBO Press have begun to publish referee reports. Publishers including Copernicus, BMJ and BMC

(the latter is owned by Springer Nature) have been doing so for even longer (see 'Revealing peer review'). Last year, the organizers of Peer Review Week embraced the topic in a broader discussion of transparency.

We are representatives of two biomedical funders — the UK Wellcome Trust and the Howard Hughes Medical Institute (HHMI)

in Chevy Chase, Maryland — and ASAPbio, a non-profit organization that encourages innovation in life-sciences publishing. We are convinced that publishing referee reports would better inform authors and readers, improve review practices and boost trust in science. Right now, less than 3% of scientific journals allow peer reviews to be ►

REVEALING PEER REVIEW

Journal editors have long consulted referees to select and improve papers. The focus has shifted to sharing them.



► published (see go.nature.com/2weh6vn).

To increase these numbers, our organizations held a meeting in February this year of around 90 invitees from the life sciences, predominantly from North America and Europe. Scientific authors, reviewers and readers participated, along with journal editors and leaders of granting agencies. We took care to include conservative voices, but the nature of the meeting attracted people ready for change. The ideas in this article were honed at that event, with later assistance from HHMI president Erin O'Shea; molecular biologist Needhi Bhalla at the University of California, Santa Cruz; Kenneth Gibbs, director of postgraduate training at the US National Institute of General Medical Sciences; and researcher Tony Ross-Hellauer at Know-Center in Graz, Austria.

Attendees agreed that the current lack of transparency around peer review does not serve science, and several journals committed to publishing reviews (although not necessarily reviewers' identities) and author rebuttals. Here, we invite more journals to take up the cause. It is time for transparency to become the norm.

DEFINING REVIEW

The term 'open review' has many interpretations. 'Open identities' means disclosing reviewers' names; 'open reports' (also called transparent reviews or published peer review) means publishing the content of reviews. Journals might offer one or the other, neither or both¹.

In a 2016 survey², 59% of 3,062 respondents were in favour of open reports. Only 31% favoured open identities, which they feared could cause reviewers to weaken their criticisms or could lead to retaliation from authors. Here, we advocate for open reports as the default and for open identities to be optional, not mandatory.

The vast majority of scientists think that peer review is essential for vetting research papers³. The process gives authors constructive feedback, offers editors insight and assures readers of the trustworthiness of research. Generally, however, only editors, authors and (sometimes) reviewers see referee reports. That enables several forms of abuse: referees might be superficial, rude or biased; authors might respond inadequately to reasonable criticism; editors might not hold authors or reviewers to account; and predatory publishers will charge fees without providing quality review.

Many benefits would accrue from publishing peer reviews (see 'Potential benefits of published review'). The scientific community would learn from reviewers' and editors' insights. Social scientists could collect data (for example, on biases among reviewers or the efficiency of error identification by reviewers) that might improve the process. Early-career researchers could learn by

example. And the public would not be asked to place its faith in hidden assessments.

Studies of published peer reviews are small and often also involve open identities or other innovations, making effects hard to ascertain. Nonetheless, evidence so far suggests that the scientific community finds published reports valuable. At *The EMBO Journal*, peer-review files receive about 10% of the hits the papers themselves do⁴. A pilot by the publisher Elsevier found that one-third of its website visitors accessed peer-review reports, and several editors said they used published reports as instructive examples for inexperienced reviewers (see go.nature.com/2oujfgv). Editors at the *European Journal of Neuroscience*, which launched transparent review at the end of 2016, report that referees are writing better reviews and returning them more promptly (see go.nature.com/2oxgtyf).

BARRIERS, PERCEIVED AND REAL

So why is the practice still rare? There are several reasons — some inertial, some conceptual.

Some disciplines are more keen than others. *Nature Communications* found that, given a choice, authors (and reviewers) of more than 70% of its evolution and ecology submissions opted for published reports. The figure was less than 50% for submissions in atomic, particle and theoretical physics⁵.

One concern is that, even if public reviews are anonymous, they might make reviewers reluctant to accept assignments or to criticize freely, because authors could resent criticism and retaliate against their presumed reviewers. *The BMJ* found that publishing peer-review reports with reviewers' names did not change the quality of the peer reviews, suggesting that reviewers were not intimidated⁶. What is more, authors read unsigned reviewers' reports during standard review anyway.

A bigger concern is that published reviews might be used unfairly in subsequent evaluation of the authors for grants, jobs, awards or promotions. There are few data about whether and how authors' ethnicity, gender, country of origin or institution affect the evaluations of papers. Yet there is evidence of bias in scientific publishing. Women, for example, are less likely to be first or last authors in high-profile journals, and are less likely to be asked for peer reviews^{7,8} (see also go.nature.com/2pzyvcw). And workplace evaluations of female professionals also show gender bias (see go.nature.com/2ppat2k). So the concern is that individuals from under-represented minorities could receive biased reviews. Assessors for funding, hiring and promotions could pay more attention to negative comments when authors are from under-represented groups or less-prestigious institutions. Some fields are also more critical or competitive, which might skew reviews.

Making referee reports open could allow more-effective research into how competition and bias affect the process. Meanwhile, anyone participating in open peer review — or evaluating it after the fact — should be aware of this potential for unfairness.

Another risk is the ‘weaponization’ of reviewer reports. Opponents of certain types of research (for example, on genetically modified organisms, climate change and vaccines) could take critical remarks in peer reviews out of context or mischaracterize disagreements to undermine public trust in the paper, the field or science as a whole. Queries to *eLife*, *The BMJ* and EMBO Press about this problem revealed only one, mild example (see go.nature.com/2piygkb). But weaponization could be a greater concern for journals that publish work that is more likely to be politicized.

One precaution would be to add a disclaimer explaining the peer-review process and its role in scientific discussion. Opening up materials and establishing dialogues with journalists, politicians and the public is an opportunity to build trust and enhance understanding of the scientific process.

Published peer-review reports could also place editorial decisions under greater scrutiny and perhaps make editors more timid about overriding critical reviews (see go.nature.com/2bid8ag). Equally, published reports could boost appreciation for the role of editors in synthesizing and prioritizing diverse reviewer opinions. Editorial judgments have huge influence on our perceptions of scientific quality, and as such are valuable works of scholarship. Publishing decision letters, along with author responses, could contextualize and rebut criticism, correct misunderstandings and provide information that does not make it into the final paper.

Finally, there are pragmatic concerns. Editors report that manually posting

peer-review materials can take approximately 25 minutes per manuscript. This is obviously much less than the time spent coordinating, conducting and assessing reviews, but is still significant. Most publishing platforms are not set up to display, organize or assign digital object identifiers

“It is time for transparency to become the norm.”

(DOIs) to reviewer reports and related materials, and making changes to such systems can be onerous. Still, we expect that journals could streamline these tasks and, potentially, build in transparent costs for dealing with extra work. Editors will have to learn to handle reviews containing inappropriate material, such as libellous comments or unpublished results, and to become comfortable with making some correspondence public. Many have already done so.

MOVING FORWARD

We think that the value of published review reports to referees, authors, the public and editors far outweighs the risks and toil. In an ideal world, all published papers would be accompanied by the contents of their peer-review reports. For now, we recommend that the practice is encouraged while the scientific community assesses whether and how author characteristics, such as ethnicity and country of origin, influence reviewer feedback. Any structural barriers to equality must be eliminated.

Ideally, reviewer reports will be easy to find, and will be organized and archived intuitively with related materials (such as rebuttals and manuscript versions). Publishers should provide ways to recognize reviewers for their contributions. Technological innovation could reduce administrative burdens, tackle information overload and provide additional links and context for readers. Vendors of

manuscript-management platforms should develop workflows that optimize and automate the process of publishing peer reviews, reducing burdens on journal staff, authors and reviewers. Indexing services, such as PubMed, should find ways to prominently link peer reviews to the original paper. Appropriate infrastructure is already being built: CrossRef began assigning DOIs for peer reviews in late 2017, and reviewer reports (with or without reviewer identities) can be archived in PubMed Central and Europe PubMed Central.

For robust systems to develop, however, published reviews must become more common. Even if today’s implementations are less than ideal, they will drive demand and pave the way for better iterations.

More than 20 editors and publishers representing more than 100 journals have already signed an open letter (see <http://asapbio.org/letter>) to show that they have begun to publish peer reviews, with or without reviewers’ identities, or that they plan to. We invite others to join. (*Nature Communications* has signed this pledge. Other *Nature* journals are considering doing so.)

Scientists can stimulate this change by requesting that the journals for which they write, review and edit are open about their peer-review process. Funders and academic societies could also help to shift attitudes, particularly if they implement or pilot published peer reviews in their society journals.

Science moves forward through criticism and disagreement. Exposing this inherent process, although uncomfortable for some, is a healthy step for science. ■

Jessica K. Polka is executive director of ASAPbio in San Francisco, California, USA. **Robert Kiley** is head of open research at the Wellcome Trust in London. **Boyana Konforti** is director of scientific strategy and development, and **Bodo Stern** is chief development and strategy officer at the Howard Hughes Medical Institute in Chevy Chase, Maryland, USA. **Ronald D. Vale** is president of ASAPbio in San Francisco, California, USA; a professor of cellular and molecular pharmacology at the University of California, San Francisco; and an HHMI investigator.

e-mails: jessica.polka@asapbio.org; ron.vale@ucsf.edu

TRANSPARENT CRITIQUE

Potential benefits of published review

- **Encourages good-quality, constructive comments.** The expectation that reviews will be published will encourage editors and reviewers to hold them to a high standard.
- **Preserves useful scholarship.** Peer reviews contain arguments and ideas that can reveal how thinking in a field evolves. This material should be preserved and made available to others.
- **Builds trust.** Readers have a right to understand the level of scrutiny that a paper has undergone.
- **Makes journal decisions more transparent.** Editors must integrate information from diverse sources, including

reviewers, to make their decisions. Published peer review provides a window on the process.

- **Creates a pathway for crediting reviewing.** Reviewers can point (even privately) to their work as evidence of scholarly activity for grants and promotions.
- **Provides a resource for training.** Reports can show people how to (and how not to) assess a paper.
- **Bolsters systemic study of peer review.** Published reports and rebuttals enable more research on best practices, leading to improvements in the system as a whole.

1. Ross-Hellauer, T. *F1000Research* **6**, 588 (2017).
2. Ross-Hellauer, T., Deppe, A. & Schmidt, B. *PLoS ONE* **12**, e0189311 (2017).
3. Publishing Research Consortium. *PRC Peer Review Survey 2015* (Mark Ware Consulting, 2016).
4. Pulverer, B. *EMBO J.* **29**, 3891–3892 (2010).
5. *Nature Commun.* **7**, 13626 (2016).
6. van Rooyen, S., Delamothe, T. & Evans, S. J. W. *Br. Med. J.* **341**, c5729 (2010).
7. Tregenza, T. *Trends Ecol. Evol.* **17**, 349–350 (2002).
8. Shen, Y. A., Webster, J. M., Shoda, Y. & Fine, I. Preprint at bioRxiv <https://doi.org/10.1101/275362> (2018).



Physicist Erwin Schrödinger also probed questions of molecular biology.

IN RETROSPECT

What Is Life?

Philip Ball revisits Erwin Schrödinger's influential book, which crystallized key concepts in modern biology.

In *What Is Life?* (1944), Austrian physicist and Nobel laureate Erwin Schrödinger used that (still-unresolved) question to frame a more specific but equally provocative one. What is it about living systems, he asked, that seems to put them at odds with the known laws of physics? The answer he offered looks prescient now: life is distinguished by a “code-script” that directs cellular organization and heredity, while apparently enabling organisms to suspend the second law of thermodynamics.

These ideas inspired the public and a number of scientific luminaries, but exasperated others. Although their elements were not original, the formulation brilliantly anticipated Francis Crick and James

What Is Life? The Physical Aspect of the Living Cell
ERWIN SCHRÖDINGER
Cambridge University Press (1944)

Watson's discovery in

Elegant and accessible, *What Is Life?* grew from a series of enormously popular public lectures that Schrödinger gave at Trinity College Dublin in 1943, in the depths of the Second World War. Exiled from Austria when it was annexed by Nazi Germany, Schrödinger had been invited to Ireland to help establish the Dublin Institute for Advanced Studies. (This September, Trinity will mark the

1953 of how DNA's double helix encodes genes. As Crick wrote to Schrödinger that year, he and Watson had “both been influ-

enced by your little book.”

lectures' anniversary with a conference called Schrödinger at 75 — The Future of Biology.) Since the 1930s, biology had been turning from a largely descriptive science into one concerned with mechanism. Thanks to studies such as those by geneticist Thomas Hunt Morgan on fruit flies, researchers were starting to understand heredity in terms of the transmission of genes, envisaged as large molecules arranged on chromosomes. Many expected genes to be proteins. However, even as Schrödinger was preparing his lectures, the microbiologist Oswald Avery was finding evidence that they were nucleic acids. Thus, *What Is Life?* dropped into a tumultuous time for science as well as for sociopolitics.

Schrödinger steps into these cross-disciplinary waters cautiously. He declares himself a “naïve physicist”, pondering how life sustains itself and transmits genetic mutations stably across generations. His work on quantum mechanics had earned him a Nobel prize in 1933, but that was hardly qualification for commenting on biology, in which Schrödinger had previously shown little interest beyond forays into the physiology of vision. Arguably, that naivety is the source of the book's strengths as well as its weaknesses.

The puzzle in the title stemmed from how physicists and chemists then thought of the molecular world, as wholly governed by statistical behaviour. In the classical molecular physics of James Clerk Maxwell and Ludwig Boltzmann, atomic motions are random (see E. Schrödinger *Nature* 153, 704–705; 1944). Precise, robust physical laws, such as those linking the temperature, pressure and volume of a gas, emerge from the average behaviour of countless atoms.

How, in that case, can a specific macroscopic outcome — a phenotype, an organism's observable inherited traits — arise from an individual genetic mutation at the molecular level? Here, perhaps, is a ghost of Schrödinger's cat, formulated in 1935, whose macroscopic life or death hinges on a single quantum event. (Mathematician Roger Penrose has said of the thought experiment that “it would not surprise me if Schrödinger had something of this issue partly in mind” when he wrote *What Is Life?*.) Looking at an inherited characteristic (such as the protruding lower jaw common among members of Europe's Habsburg dynasty), Schrödinger asks how the allele responsible remained “unperturbed by the disordering tendency of the heat motion for centuries?”

Here, he cites experiments by another former quantum physicist, Max Delbrück, whose use of high-energy radiation to induce genetic mutations allowed him to estimate a gene's size at around 1,000 atoms. Schrödinger claims that this seems too small for “lawful activity” — durable inheritance — to persist in the face of statistical fluctuations. But he asserts that quantum mechanics can explain the matter. Atoms

in molecules can typically be arranged in many stable ways, and each configuration has an associated energy; this is how Schrödinger envisages different gene alleles. But “quantum jumps” between them are generally inhibited by high energy barriers.

He goes on to propose that such gene-encoding molecules (he was among those who suspected that they were large proteins) have enough potential variety in their configurations to encode huge amounts of information, and that this variety can furnish a cell’s “code-script”. The position of each atom matters, but the pattern does not repeat — hence his description of the

molecules as being like an aperiodic (irregular) solid. It wasn’t an entirely new idea; Delbrück had suggested something of the kind in 1935.

And biologists

Hermann Muller and J. B. S. Haldane had independently proposed that chromosomes might act as templates for their own replication, in the same way that new crystal layers build up on pre-existing ones.

None of this, Schrödinger admits, answers the deeper question of “how the hereditary substance works” — that is, how it is used in development and metabolism, enabling an organism to build and sustain itself from moment to moment in what Schrödinger calls its “four-dimensional pattern” in space and time. But he makes a start on that issue by posing the question in thermodynamic terms.

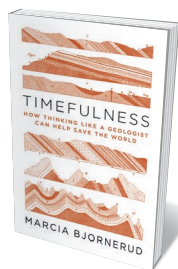
This isn’t a matter of energy (organisms’ energy intake and output must be balanced, or they’d burn up), but of entropy, the measure of atomic disorder. The second law of thermodynamics states that entropy must increase in all processes of change. But organisms somehow stave off entropic dissolution. As Schrödinger put it, they feed on “negative entropy”, using it to sustain the organization apparent in the structures and functions of cells, while paying their thermodynamic dues by heating the environment.

How they mine negative entropy, he could not say. He was forced to suggest that, in living systems, “we must be prepared to find a new type of physical law”. Today, no such drastic solution seems to be needed.

The concept missing from his analysis is information. The information theory of Claude Shannon and the cybernetics of Norbert Wiener in the 1940s and 1950s began to fill that lacuna, although only more recently have researchers begun to understand how information truly features in biology. As Schrödinger’s talk of negative entropy hinted, life is a pocket of out-of-equilibrium order in an open system, and the DNA code is just part of what sustains it. It’s a shame that Schrödinger didn’t touch on ►

“What Is Life? dropped into a tumultuous time for science as well as for sociopolitics.”

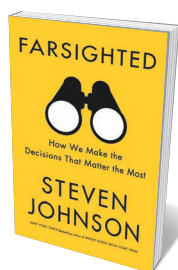
Books in brief



Timefulness

Marcia Bjornerud PRINCETON UNIVERSITY PRESS (2018)

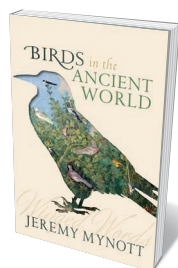
As a geologist, Marcia Bjornerud works in many time frames: the 4.5-billion-year history of Earth, the academic year, the daily grind. That layered perspective has made her aware of the short-term thinking common in a society wedded to political terms of office and the news cycle — all of which has, she argues, contributed to our inadequate, sometimes wrongheaded response to climate change. In this trenchant study, Bjornerud calls for a new geological literacy to instil deeper knowledge of planetary rhythms and processes — “thinking like a mountain”, as ecologist Aldo Leopold put it.



Farsighted

Steven Johnson RIVERHEAD (2018)

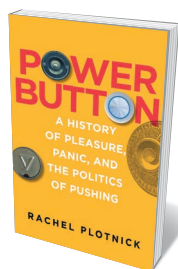
Many researchers (notably psychologist Daniel Kahneman) have wrestled with the subtle mechanics of decision-making. Now, science writer Steven Johnson has his decisive moment, looking at the deep deliberations — mapping of variables, predictions of outcomes and balancing of aims and possibilities — that underpin life-changing choices. He draws on research and compelling examples, from George Eliot’s 1871 novel *Middlemarch* (which examines the “threadlike pressure” on the deciding mind) to the supercomputer-based climate models now influencing climate-relevant decisions across the globe.



Birds in the Ancient World

Jeremy Mynott OXFORD UNIVERSITY PRESS (2018)

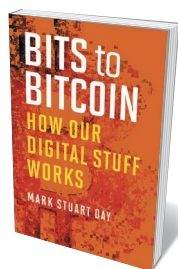
From nightingales trilling in ancient Rome’s suburbs to the migrating cranes minutely observed by Aristotle in his fourth-century-BC *History of Animals*, birds pervaded early Mediterranean civilizations. Jeremy Mynott’s masterful cultural and scientific history tours their roles as timepieces, soundscapes, pets, messaging services — even intermediaries with the supernatural. The vivid artworks and literary passages give this wings: here is the Greek poet Aratus on finches “chirruping shrilly at dawn” before a storm; there, a surreal Roman recipe for flamingo stewed with coriander.



Power Button

Rachel Plotnick MIT PRESS (2018)

Push buttons pop up on everything from blenders to aeroplanes. Yet, as Rachel Plotnick reveals in this unusual technological history, the mechanism had an explosive impact on culture from its debut in the 1880s to the 1920s and beyond. The idea that huge machines or even bombs could be activated by a finger became a metaphor for human hegemony, and a source of fear and wonder. And, as Plotnick notes, some ‘buttonized’ inventions (such as the electrified tie pin) may be defunct, but in an era of nuclear weaponry and disruptive leadership, one-touch technology still has the power to shock.



Bits to Bitcoin

Mark Stuart Day MIT PRESS (2018)

In this methodical primer, technologist Mark Day examines the computational infrastructure — the elements that underlie the workings of digital devices and networks. He unpicks operating systems, examines processes, explains esoteric defensive techniques such as cryptography and reveals Bitcoin to be an “intriguing combination of self-interest and mathematics”. If you want to know why data streams turn lumpy when compressed, or yearn to get inside the cloud, a handy reference awaits. [Barbara Kiser](#)

► fellow physicist Leo Szilard's work on Maxwell's demon, a thought experiment that revealed how entropic disorder could be undone by making use of molecular-level information that looks like mere statistical noise at the macroscopic level.

What's more, Schrödinger gave his code-script too much agency by imagining that its readout was mapped directly onto the phenotype. This isn't how it works: you can't read the arrangement of the body's organs in the genome. The information functions as a resource, not a step-by-step guide. To acquire meaning, it must have context: a cell's history and environment. Tracing how the phenotype emerges from interactions of genes with each other and with their environment is the key puzzle of modern genomics.

What is Life? helped to make influential biologists out of several physicists: Crick, Seymour Benzer and Maurice Wilkins, among others. But there's no indication from contemporary reviews that many biologists grasped the real significance of Schrödinger's code-script as a kind of active program for the organism. Some in the emerging science of molecular biology were critical. Linus Pauling and Max Perutz were both damning about the book in 1987, on the centenary of Schrödinger's birth. Pauling considered negative entropy a "negative contribution" to biology, and castigated Schrödinger for a "vague and superficial" treatment of life's thermodynamics. Perutz grumbled that "what was true in his book was not original, and most of what was original was known not to be true even when the book was written".

Although these judgements are uncharitable, they are not without substance. Why, then, was the book so influential? Rhetorical theorist Leah Ceccarelli argues that it was down to Schrödinger's writing style: he managed to bridge physics and biology without privileging either. But today, we can find more than that. Schrödinger's thoughts on the entropic balance of life can be regarded as precursors to studies of how biological prerogatives such as replication, memory, ageing, epigenetic modification and self-regulation must be understood as processes of non-equilibrium complexity that cannot ignore the environment. It is intriguing that similar considerations of environment and contingency are now seen to be central in quantum mechanics, with its ideas of entanglement, decoherence and contextuality. Whether this is more than coincidence, we can't yet say. ■

Philip Ball is a writer based in London. His latest book, on quantum physics, is *Beyond Weird*.
e-mail: p.ball@btinternet.com

SECURITY

How smart connectivity is stupider by design

Steven Aftergood assesses a warning about the future of the Internet.

Hardly a day now passes without reports of a massive breach of computer security and the theft or compromise of confidential data. That digital nightmare is about to get much worse, asserts security technologist Bruce Schneier in *Click Here to Kill Everybody*, his critique of government inertia on Internet security.

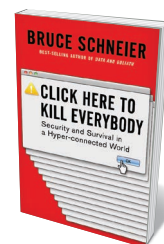
The burgeoning threat, writes Schneier, arises from the rapid expansion of online connectivity to billions of unsecured nodes. The Internet of Things, in which physical objects and devices are networked together, is well on its way to becoming an Internet of Everything. Over the past decade or so, a growing number of products have been sold with embedded software and communications capacity: household appliances, cars, medical instruments and even clothing can now be monitored and controlled from afar. More of the same is on the way, as smart homes yield to smart cities and automated systems assume a larger role in the management of critical infrastructure. The Stuxnet computer worm used to attack Iran's uranium-enrichment programme remotely in 2010 was an early, audacious indicator of the threat.

Enhanced global connectivity has many advantages for knowledge sharing, commerce and convenience. Securing it, however, is a daunting prospect. The all-too-familiar vulnerability of computer networks — their susceptibility to failure, disruption and interference by malware, viruses and other factors — is amplified as practically everything becomes computerized. That relentless expansion of cyberspace into the physical domain brings with it new threats to power systems, mass transportation, public health and safety, and even political institutions, as effectively demonstrated by the Russian information operations that targeted the 2016 US presidential election.

Despite its lurid title, Schneier's book is sober, lucid and often wise in diagnosing how the security challenges posed by the expanding Internet came about, and

in proposing what should (but probably won't) be done about them.

As he notes, security was not a primary concern in the early design of the Internet in the mid to late twentieth century. Developers of early efforts, from the US Department of Defense's ARPANET onwards, did not anticipate the Internet's explosive growth or coming



Click Here to Kill Everybody: Security and Survival in a Hyper-connected World
BRUCE SCHNEIER
W. W. Norton (2018)

role in global commerce and communication. Even today, there is little incentive to prioritize security above other concerns — so, for example, e-mails may or may not be from the sender named.

SURVEILLANCE CAPITALISM

Surprisingly to some, much of the business of the Internet is predicated on insecurity. 'Surveillance capitalism' — the collection of user data and its sale to advertisers and others — depends on vulnerable Internet practices, as does intelligence collection for national security and law enforcement. Governments act as if their need to monitor the Internet can be satisfied without any larger compromise of security. That, writes Schneier, is not so.

In principle, he explains, securing the Internet is straightforward, but it would demand concerted government action at each step. Financial incentives should be realigned to promote security and penalize failure by mandating that manufacturers disclose defects in commercial software, making them legally liable for defects. Security should be required in new devices, and rewarded through subsidies and tax breaks. Data should be encrypted to secure them against unwanted collection. Critical infrastructure — power grids, communications and transportation — should be protected by bolstering network security or disconnecting them from the network altogether.

Government agencies are fully aware that the expanding Internet "will create

"When the Internet starts killing people it will be regulated."



Airline systems, like all networks, are vulnerable to hacking.

MICHAEL H/GETTY

an incalculably larger exploitation space for cyber threat actors”, as the US National Counterintelligence and Security Center noted in a 2018 report, *Foreign Economic Espionage in Cyberspace*. Yet Schneier’s views on security differ sharply from those of many government officials in the United States and elsewhere. For instance, Schneier considers strong encryption to be indispensable for personal and network security. The US Department of Justice sees it as “a serious challenge to effective law enforcement”.

Similarly, Schneier advocates ruling out ‘back doors’ — design features that enable users, authorized or not, to bypass security and to decipher encrypted communications. He reasons that they render entire systems more vulnerable. But as then-UK home secretary Amber Rudd said last year, lack of access to encrypted data “in specific and targeted instances is right now severely limiting our agencies’ ability to stop terrorist attacks and bring criminals to justice”. Schneier also feels that it would be unfeasible and inappropriate to ban anonymity online. But the US Department of Justice insists that

impenetrable anonymity “poses a unique and significant threat to public safety” in criminal contexts.

PRIORITY CLASH

Because Schneier and his opponents in law-enforcement agencies are responding to different problems on different timescales — solving a crime today versus fixing the whole Internet for the foreseeable future — it is difficult to say categorically that one side is right and the other wrong. But Schneier argues his position well. And to compensate for the admitted loss of collection capability that would follow from improved Internet security, he proposes to “make law enforcement smarter” through security research, enhanced computer forensics and new career paths.

Although cybersecurity is a hot-button issue in policy circles, progress is hindered by bureaucratic lethargy, especially on fundamental questions. In July, the US Government Accountability Office reported that 1,000 of its recommendations for addressing cyber threats have yet to be implemented, placing government information systems

increasingly at risk. Governance seems to be an even harder problem than cybersecurity, leaving Schneier to predict that the United States “will do nothing soon”.

At some point action will become imperative — perhaps sooner in the European Union, which has demonstrated a willingness to act on data-protection issues. “Governments regulate things that kill people,” Schneier notes, citing vehicles, airlines and power plants. He adds: “when the Internet starts killing people it will be regulated”.

Not just any regulations will do. To help devise a sensible response, he says, scientists and engineers need to get more involved in the policy process. And the challenges posed by the advancing online world go beyond security. If the question is what sort of Internet is compatible with a humane and enlightened society, technologists are not the only ones who will need a seat at the table. ■

Steven Aftergood directs the *Federation of American Scientists Project on Government Secrecy in Washington DC*.
e-mail: saftergood@fas.org

Correspondence

Preprints: safeguard rigour together

Tom Sheldon's concern that preprints might lead to poor research being overblown in the media is more likely to apply to the press releases circulated to journalists under embargo than to the preprints themselves (*Nature* 559, 445; 2018). Wherever they hear about a story, journalists are under the same obligation as scientists to critically review the work they intend to communicate to readers.

When journalists try to secure independent expert opinions, they should indicate whether and how preprint manuscripts have been screened — in keeping with disclaimers on some preprint servers. And scientists can impede the spread of low-quality information by publicly commenting on preprints and peer-reviewed papers, giving readers an insight into the scientific community's reaction to a work.

The increasing popularity of preprints is an opportunity for researchers, institutions, funders and journalists to coordinate discussion of how research is covered in the media.

James Fraser *University of California, San Francisco, USA.*

Jessica Polka* *ASAPbio, San Francisco, California, USA.*
J.P. declares competing interests (see go.nature.com/2wnffew).
jessica.polka@asapbio.org

Preprints: good for science and public

We disagree with Tom Sheldon's contention that the preprint ecosystem can present a challenge to accurate and timely journalism (*Nature* 559, 445; 2018). Restricting when or how preprints are released risks suppressing science communication without any clear advantage to the public.

When scientists and journalists follow fundamental principles for reporting

research results — such as ensuring that publications are rigorously sourced and fact-checked — preprints pose no greater risk to the public's understanding of science than do peer-reviewed articles (S. Sarabipour *et al.* *PeerJ Preprints* 6, e27098v1; 2018).

Responsible journalists already report on preprints with the help of real-time commentary from scientists on Twitter and elsewhere (see go.nature.com/2kctmfn). Peer-reviewed papers are published under an embargo, so this important resource is not available.

Preprints lead to scientific collaborations, reagent requests and adoption of new techniques. And as scientists benefit increasingly from preprints and other pre-publication research outputs, so too will the public. **Sarvenaz Sarabipour*** *Johns Hopkins University, Baltimore, Maryland, USA.*
**On behalf of 9 co-signatories; competing interests declared (see go.nature.com/2p9gqwm for details).*
ssarabi2@jhu.edu

Preprints: help not hinder journalism

In suggesting that preprints could distort the public's understanding of science, Tom Sheldon perpetuates the fallacy that peer review is a guarantee of validity (*Nature* 559, 445; 2018). There are countless examples to the contrary (see, for instance, A. Margalida and M. À. Colomer *PeerJ* 4, e1670; 2016).

A responsible journalist consults multiple independent sources to verify research findings. This critical evaluation is not contingent on the research having been peer reviewed. Preprints provide early and unrestricted dissemination of research outputs, so journalists can often peruse expert feedback when considering a story. And most preprint servers either label

preprints as 'not peer reviewed' or have editorial 'sanity checks' in place to prevent the posting of junk science.

Plenty of peer-reviewed research papers contain errors. Preprints provide a chance to spot these and have them removed before publication. In our view, preprints and peer review are complementary.

Jonathan Tennant* *Open Science MOOC, Leicester, UK.*
Laurent Gatto *de Duve Institute, Catholic University of Louvain, Brussels, Belgium.*
Corina Logan *Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.*
**J.T. declares competing interests (see go.nature.com/2o4klgk).*
jon.tennant.2@gmail.com

Border carbon fees could rebound

We agree with Michael Mehling and colleagues that applying carbon charges — rather than trade tariffs — to imports could help to address countries' non-compliance with climate policy (*Nature* 559, 321–324; 2018). However, their advice to match these charges (known as border carbon adjustments) to the cost of domestic carbon is economically questionable.

Although such charges would level the playing field for domestic and external manufacturers, the same is not true for consumers, domestic or external. The fees would make carbon-intensive goods cheaper for consumers in unregulated countries, and so boost consumption. And they would reduce US exports of ferrous metal products to the European Union, say, while increasing the supply and lowering the price of steel products in the United States.

This consumption-rebound effect could mean there is a smaller drop in carbon emissions than would be expected from imposing border carbon-adjustment charges. Charges motivated purely by

climate considerations would therefore need to be below the domestic cost. If set at the domestic level, they could be a form of protectionism: the country levying the border carbon charges would benefit from its trade power (above and beyond climate management) at the expense of the nation targeted.

Such economic complexities indicate that border carbon adjustments are an imperfect substitute for negotiating international agreements on carbon emissions. **Edward Balistreri** *Iowa State University, Ames, Iowa, USA.*
Daniel Kaffine *University of Colorado Boulder, USA.*
Hidemichi Yonezawa *Statistics Norway, Oslo, Norway.*
daniel.kaffine@colorado.edu

Never mind the gold watch

You note that some universities grant emeritus status only to those professors who have a distinguished research record, whereas others automatically bestow the honour on all retiring full professors (*Nature* 559, 429–431; 2018). As an emeritus professor, I would like to point out that emeritus — an unusual word in that it is derived from two classical roots, rather than one — holds as well for both: 'e' is from the Greek for out, and 'meritus' from the Latin for 'deserving, meritorious', or, more loosely, you deserve to be.

David Rickard *Cardiff University, UK.*
rickard@cardiff.ac.uk

CONTRIBUTIONS

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at <http://go.nature.com/cmchno>.

Burton Richter

(1931–2018)

Physicist who helped to discover the first particle containing a charm quark.

The burst of scientific activity that began with the discovery of the J/ψ particle in 1974 is known to particle physicists as the ‘November revolution’, because it so radically changed their perspective. Charm quarks, heavier than the quarks that make up protons and neutrons, were predicted by theory that became the standard model of particle physics. Finding particles containing them opened up a new chapter in physics.

Burton Richter, who died on 18 July, made that revolution possible by designing and building the positron–electron accelerator SPEAR and an innovative detector facility, both at the Stanford Linear Accelerator Center (SLAC) in California.

Richter received the 1976 Nobel Prize in Physics for the J/ψ discovery. (The double name for the particle is because two experimental groups announced their results on the same day; Samuel Ting, who led the other experiment, shared the Nobel prize.)

It has to be said that Burt was not looking for this particle. He agreed to let his colleagues, Marty Breidenbach, Vera Lüth and Roy Schwitters, “waste a weekend” (as he put it) to redo measurements to check an anomaly in their data. They found this new particle immediately — noting in the log book that the data “came pouring in”. Within months, they made further discoveries that confirmed their interpretation of the particle as a charm quark.

This breakthrough would not have happened without Richter’s energy and persistence. Most physicists did not see SPEAR as a priority, and the US Congress would not fund it. After years of rejections, Richter redesigned the project to cut costs. He convinced his laboratory director and the Atomic Energy Commission to let him build it using ongoing funding. Once repackaged as an improvement to an existing facility, rather than a new project, it did not need congressional approval.

Burton Richter was born in Brooklyn, New York, on 22 March 1931, and grew up in Queens, another borough of the city. Graduating from high school with a passion for science experiments, he enrolled at the Massachusetts Institute of Technology, Cambridge, where he obtained his undergraduate degree in 1952 and his PhD in physics four years later. He moved to Stanford University and in 1963 to the new Stanford Linear Accelerator Center,



where he spent the rest of his career. He was director of SLAC from 1984 until 1999, and remained involved in physics research and policy until the day he died.

Particle physicists generally fall into one of three categories: theorist, experimentalist or accelerator physicist. Each requires specialized expertise, and so few people master more than one. Burton Richter was the exception. He excelled at designing and building accelerators, as well as designing and leading experiments that used them.

Richter’s enthusiasm for storage rings — accelerators in which particles circulate for hours — began at Stanford. With physicist Gerald O’Neill of Princeton University, in New Jersey, Richter helped to build the world’s first pair of rings to store electrons. SPEAR was his next project, with electrons and positrons held in a single ring. The positron–electron project (PEP), a larger and higher-energy storage ring, came online at SLAC in 1980.

Burt was keen to push for ever-higher particle energies. To produce lots of Z-particles, a carrier of the weak nuclear force, he invented the SLAC linear collider. Short pulses of high-energy electrons and positrons, in beams the width of a human hair, travel for miles along the SLAC accelerator. They pass along separate arcs and then collide, occasionally producing Z-particles. The collider yielded, among other results, a tight upper bound on the mass of the Higgs boson and, just as importantly, it demonstrated the viability of the linear-collider concept for future

machines. To achieve even-higher-energy collisions, Burt long advocated setting up two linear accelerators head-to-head. He instigated efforts to plan one, but such a machine has yet to be built.

Richter recognized that SLAC had to diversify to survive. When SPEAR was under construction, condensed-matter physicists Sebastian Doniach and William Spicer at Stanford University convinced him to add a small window in the vacuum pipe of the storage ring, so that the X-rays produced by circulating electrons or positrons could get out. This created the most intense X-ray source then available, opening the door to X-ray science. Intense X-ray pulses have many applications in materials science, chemistry and in deciphering biological structures. Today, this science is one of the main uses of SLAC.

Richter also brought particle astrophysics to the science mix at the lab, and supported a plan to convert the SLAC accelerator into the world’s first X-ray laser (itself a tour de force of accelerator design) to produce pulses of even-more-intense X-rays.

Burt served the community in other ways. He was a member of the JASON group, providing technical advice to the US government. He served on countless national and international science-advisory panels, and as councillor and president of the American Physical Society. His interest in energy policy led to his encyclopaedic book *Beyond Smoke and Mirrors: Climate Change and Energy in the 21st Century*.

Burt Richter cared about ideas, not status or recognition. He would share his technical insight and acumen with anyone to improve an experiment. If a postdoc or student had a good idea, he would support it; if he disagreed with a high-status scientist, he made it clear. He argued furiously in the service of getting to the answers, and expected others to do the same. He gave responsibility to those he found most capable, without regard to seniority.

Burt loved doing physics, constructing experiments capable of reaching frontiers and breaking beyond them. ■

Helen Quinn is a theoretical physicist and professor emerita at SLAC, where she did her PhD and spent most of her career as a staff scientist and faculty member.
e-mail: quinn@slac.stanford.edu

EDDIE ADAMS/AP/REX/SHUTTERSTOCK



MARTY MELVILLE/AFP/GETTY

Figure 1 | Damage caused by the 2011 Christchurch earthquake. Large earthquakes can be followed by thousands of smaller ones, called aftershocks. In February 2011, an aftershock struck the city of Christchurch, New Zealand, and was more destructive than the earthquake it followed. The image shows the smoking ruins of the six-storey Canterbury Television building, which collapsed and caught fire in the aftershock, killing more than 100 people.

GEOPHYSICS

Aftershock forecasts turn to AI

Understanding how earthquakes interact is key to reliable earthquake forecasting. A machine-learning study reveals how the stress change induced by earthquakes at geological faults affects these interactions. [SEE LETTER P.632](#)

GREGORY C. BEROZA

All major earthquakes are followed by smaller ones, called aftershocks, which can themselves be hazardous. The forecasting of aftershocks is an area of long-standing seismological interest¹. It is receiving renewed attention² because of earthquake sequences in Italy, New Zealand and Japan over the past decade, in which the first earthquake in the sequence was not the most destructive (Fig. 1). On page 632, DeVries *et al.*³ use machine-learning tools to take a

fresh look at how changes in geological stress generated by earthquakes influence the spatial distribution of aftershocks. The authors' work provides more-accurate forecasts of aftershock locations than does the standard approach⁴.

Deterministic earthquake prediction remains an elusive goal, but seismologists are working intently to make quantitative probabilistic forecasts of future earthquake occurrences. Prominent among the factors thought to affect earthquake probabilities is the change in stress induced by one earthquake at the potential initiation site of another.

Probabilistic forecasting depends on well-established statistical properties of seismicity — the spatial and temporal distribution of earthquakes. Although earthquakes cluster in space and time, large ones are rare, which makes documenting the interactions between these earthquakes intrinsically challenging.

Large earthquakes, however, can be followed by thousands of aftershocks, which are indistinguishable from other earthquakes. Aftershocks occur by the same mechanism, on the same geological faults and under the same conditions as for other earthquakes. It is therefore

reasonable to assume that understanding the interactions between the largest earthquake in a sequence (the mainshock) and its aftershocks will enhance general understanding of earthquake interactions.

DeVries and colleagues studied these mainshock–aftershock interactions using a database of published distributions of mainshock-induced slip — the relative movement of geological features on opposite sides of a fault. From these distributions, the authors calculated the stress changes induced by the mainshocks. They fed this information into an artificial-intelligence system known as an artificial neural network, which was trained to determine the likelihood that aftershocks would occur in a particular location on a spatial grid.

The authors withheld a randomly selected 25% of the mainshock–aftershock sequences from the training data, and used this subset to validate the predictive power of their machine-learning method. They report that the trained network can predict the locations of aftershocks more accurately than can the standard forecasting approach, which considers only one aspect of the induced shift in stress, known as the Coulomb failure stress change⁴. The authors find that other characteristics of the stress change play a crucial part in triggering aftershocks. The paper therefore demonstrates how machine learning could aid research in seismology.

However, for several reasons, it might be premature to infer that DeVries and colleagues' work has led to an improved physical understanding of aftershock triggering. One reason is that the current study — and earlier studies of aftershock triggering — focused on static stress changes that occur and persist long after the passage of seismic waves⁵. But dynamic stress changes caused by seismic waves can also trigger earthquakes⁶. The combination of static and dynamic stress changes leads to a spatial distribution of aftershocks that differs from the pattern caused by static stress changes alone⁷.

Another reason for caution is that the authors' analysis relies on factors that are fraught with uncertainty. Uncertainties in earthquake locations are probably small, but uncertainties in slip distributions, on which the stress-change analysis depends, are large and potentially problematic. It is well documented that estimates of slip made by different investigators are subject to substantial differences⁸. The inferred stress change, which is input to the authors' machine-learning algorithm, depends on the rate of change of these slip distributions with respect to position, such that slip uncertainty is amplified. This issue is more problematic close to the fault than it is farther away from it, but most aftershocks occur close to the fault.

The situation is compounded by the fact that slip estimates invariably assume that slip occurs on faults that are planar or composed of multiple planes. However, fault geometry is known to be complex at all scales⁹.

This complexity leads to strong, local stress concentrations that can trigger aftershocks¹⁰, but that will not be included in slip models that assume planar faults. This could explain why the authors see no evidence of a lack of aftershocks near faults — caused by an overall decrease in stress — despite the fact that this feature is readily apparent in situations in which data and circumstances allow it to be clearly observed¹¹. These issues concerning uncertainty are not particular to the authors' study, but they counsel some temperance in calling for new physical models to explain the current results.

Regardless of the physical interpretation, the performance of DeVries and colleagues' artificial neural network is motivating. Until a few years ago, most statistical forecasts of aftershocks were more accurate than were physics-based forecasts, such as that of the authors. But there are now cases in which physics-based forecasting performs as well as purely statistical approaches^{12,13}. The time would seem ripe for methods based on artificial intelligence to enter the fray, and the work of DeVries *et al.* has established this beachhead.

Artificial-intelligence methods have much to offer seismology, and solid-Earth science more broadly. There are societally important phenomena to understand that are informed

by data sets growing rapidly in scale and scope, and by computational simulations growing rapidly in sophistication and realism. The application of machine-learning methods has the potential to extract meaning from these large and complex sources of information, but we are still in the early stages of this process. ■

Gregory C. Beroza is in the Department of Geophysics, Stanford University, Stanford, California 94305-2215, USA.
e-mail: beroza@stanford.edu

1. Reasenberg, P. A. & Jones, L. M. *Science* **243**, 1173–1176 (1989).
2. Marzocchi, W., Taroni, M. & Falcone, G. *Sci. Adv.* **3**, e1701239 (2017).
3. DeVries, P. M. R., Viégas, F., Wattenberg, M. & Meade, B. J. *Nature* **560**, 632–634 (2018).
4. King, G. C., Stein, R. S. & Lin, J. *Bull. Seismol. Soc. Am.* **84**, 935–953 (1994).
5. Stein, R. S. *Nature* **402**, 605–609 (1999).
6. Hill, D. P. *et al. Science* **260**, 1617–1623 (1993).
7. Meng, X. & Peng, Z. *Geophys. J. Int.* **197**, 1750–1762 (2014).
8. Mai, P. M. *et al. Seismol. Res. Lett.* **87**, 690–708 (2016).
9. Tchalenko, J. S. *GSA Bull.* **81**, 1625–1640 (1970).
10. Smith, D. E. & Dieterich, J. H. *Pure Appl. Geophys.* **167**, 1067–1085 (2010).
11. Toda, S., Stein, R. S., Beroza, G. C. & Marsan, D. *Nature Geosci.* **5**, 410–413 (2012).
12. Segou, M. & Parsons, T. *Seismol. Res. Lett.* **87**, 816–825 (2016).
13. Cattania, C. *et al. Seismol. Res. Lett.* **89**, 1238–1250 (2018).

PRECISION MEDICINE

Sequence of events in prostate cancer

Whole-genome sequencing reveals the duplication of a regulatory region, called an enhancer, of the AR gene in treatment-resistant human prostate cancers. The finding shows the importance of analysing non-protein-coding regions of DNA.

KELLIE A. COTTER & MARK A. RUBIN

The publication^{1,2} of the human genome sequence in 2001 was accompanied by optimism that a rise in the availability of genomic data might improve clinical treatments. It was hoped that such data might one day enable an approach termed 'precision medicine', in which therapies are tailored to target the abnormalities specific to a particular cancer. Since then, technological advances in DNA-sequencing techniques, combined with substantially lower costs, have led to a boom in the sequencing of cancer samples. Given this progress, one might assume that the key genetic alterations that drive common cancers are already well known. However, writing in *Cell*, Takeda *et al.*³, Viswanathan *et al.*⁴ and Quigley *et al.*⁵ detail a previously unidentified type of genetic alteration that frequently occurs in late-stage human prostate cancer.

The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have undertaken some of the largest-scale projects reported so far to sequence the DNA of human cancers. These efforts have identified many DNA alterations that drive cancer growth, including mutations and genomic rearrangements. TCGA has sequenced the protein-coding regions of approximately 11,000 individual genomes and 33 types of cancer (<https://portal.gdc.cancer.gov>), whereas the ICGC has sequenced the protein-coding regions from more than 20,000 individual genomes and 22 kinds of cancer (<https://dcc.icgc.org>). Both projects have focused mainly on sequencing the protein-coding regions of genes, which represent less than 2% of the entire genome. In the Pan Cancer Analysis of Whole Genomes (PCAWG) project, the ICGC and TCGA systematically analysed whole-genome

PRECISION MEDICINE

Sequence of events in prostate cancer

Whole-genome sequencing reveals the duplication of a regulatory region, called an enhancer, of the *AR* gene in treatment-resistant human prostate cancers. The finding shows the importance of analysing non-protein-coding regions of DNA.

KELLIE A. COTTER & MARK A. RUBIN

The publication^{1,2} of the human genome sequence in 2001 was accompanied by optimism that a rise in the availability of genomic data might improve clinical treatments. It was hoped that such data might one day enable an approach termed ‘precision medicine’, in which therapies are tailored to target the abnormalities specific to a particular cancer. Since then, technological advances in DNA-sequencing techniques, combined with substantially lower costs, have led to a boom in the sequencing of cancer samples. Given this progress, one might assume that the key genetic alterations that drive common cancers are already well known. However, writing in *Cell*, Takeda *et al.*³, Viswanathan *et al.*⁴ and Quigley *et al.*⁵ detail a previously unidentified type of genetic alteration that frequently occurs in late-stage human prostate cancer.

The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have undertaken some of the largest-scale projects reported so far to sequence the DNA of human cancers. These efforts have identified many DNA alterations that drive cancer growth, including mutations and genomic rearrangements. TCGA has sequenced the protein-coding regions of approximately 11,000 individual genomes and 33 types of cancer (<https://portal.gdc.cancer.gov>), whereas the ICGC has sequenced the protein-coding regions from more than 20,000 individual genomes and 22 kinds of cancer (<https://dcc.icgc.org>). Both projects have focused mainly on sequencing the protein-coding regions of genes, which represent less than 2% of the entire genome. In the Pan Cancer Analysis of Whole Genomes (PCAWG) project, the ICGC and TCGA systematically analysed whole-genome sequencing data from many types of cancer. These data allowed scientists to investigate alterations in DNA regions that regulate gene expression, and in untranslated parts of gene sequences. This revealed that, in cancer cells, alterations in these non-protein-coding regions of DNA

occur at a similar frequency to those in the protein-coding regions⁶.

Many of the sequencing studies reported by TCGA and the ICGC focused predominantly on tumour samples taken from patients before cancer treatment. The work of Takeda, Viswanathan, Quigley and their respective colleagues provides some information about genetic alterations present in prostate cancers that are resistant to clinical treatment.

Prostate-cancer growth is usually driven by signalling pathways that act through the androgen receptor (AR), and a standard clinical treatment for advanced prostate cancer is to reduce the level of androgen hormones that activate ARs. Although this limits cancer growth for a while, tumours eventually become resistant to this therapy, and a highly malignant form of the cancer arises that is usually lethal. Such a tumour can migrate to other sites in the body through a process known as metastasis, and this sort of late-stage, treatment-resistant tumour is called a metastatic, castration-resistant prostate cancer.

When treatment resistance occurs, an

altered version of the gene that encodes AR is commonly found in the tumour. Mutations of the *AR* gene or amplifications of DNA that increase the copies of sequence encoding *AR* might enable tumour cells to enhance AR-pathway signalling even when androgen levels are low^{7,8}. Analyses of protein-coding-sequence changes linked to prostate cancer have found alterations in *AR*, as well as in other known cancer-promoting genes⁹. Although a wealth of DNA sequencing data of protein-coding regions are available for prostate-cancer samples, there are comparatively few whole-genome sequences (only approximately 200 have been reported by the PCAWG project, for example; <https://dcc.icgc.org/pcawg>), and still fewer whole-genome sequencing data are available for metastatic, castration-resistant prostate cancer.

Takeda *et al.* re-evaluated previously published data¹⁰ from clinical samples of castration-resistant prostate cancer and identified repeated DNA sequences that caused abnormal amplification of the region upstream of *AR* (Fig. 1). The authors describe this region as a type of gene-regulatory element called an enhancer, which is a sequence that can help to promote gene expression. When Takeda and colleagues used a genome-editing technique to target and suppress this region in human prostate-cancer cells grown *in vitro*, both cell proliferation and AR expression were reduced. The authors also engineered prostate-cancer cells grown *in vitro* to contain a duplication of the enhancer, and found that such cells showed increased AR expression and decreased sensitivity to an AR-targeting drug called enzalutamide that is used to treat metastatic, castration-resistant prostate cancer.

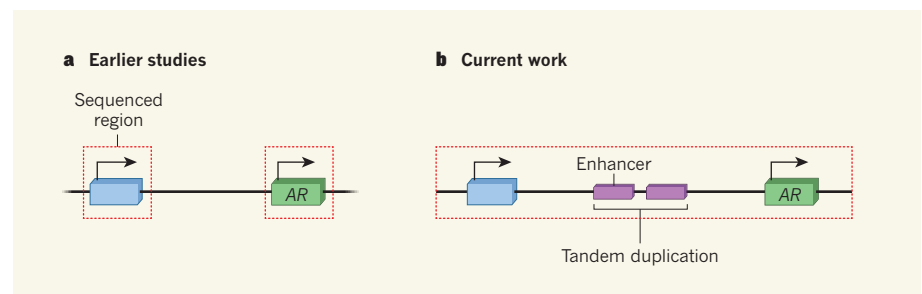


Figure 1 | Duplication of an enhancer region of the genome occurs frequently in prostate cancer. **a**, Many of the earlier DNA-sequencing studies of human prostate cancer have focused on the protein-coding regions of the genome, such as the gene shown in blue (the red dotted box indicates sequenced regions). This work identified alterations in the sequence of the *AR* gene, which encodes the androgen receptor (AR), as being a common driver of disease progression^{7,8}. **b**, Takeda *et al.*³, Viswanathan *et al.*⁴ and Quigley *et al.*⁵ demonstrate the utility of sequencing approaches that are not restricted to the protein-coding regions, and the advantages of sequencing tumour samples that have become resistant to therapy as a way of investigating why clinical treatment eventually fails. The three studies of late-stage prostate cancer report that the DNA sequence in a region upstream of *AR*, termed an enhancer, is commonly expanded, and this amplification is often in the form of a tandem duplication of the sequence. An enhancer amplification can drive expression of *AR*, which would enable tumours to evade the effects of clinical treatments that target the AR signalling pathway.

Viswanathan *et al.* present whole-genome sequencing data for 23 samples of metastatic, castration-resistant prostate cancer from patients. The authors compared this sequencing data with matched data from non-cancer cells from these individuals. This enabled the researchers to report alterations that characterize this type of prostate cancer. These included numerous duplicated sequences, with many duplications occurring in a tandem pattern. These tandem duplications frequently occurred in genome sequences adjacent to *AR* and to another cancer-promoting gene called *MYC*. The main region of sequence amplification associated with *AR* was in the same enhancer region that was identified by Takeda and colleagues. Viswanathan and colleagues found that the enhancer amplification was present in 87% of their samples, either with or without an amplified copy of *AR*.

Quigley and colleagues performed whole-genome sequencing of 101 samples of metastatic, castration-resistant prostate-cancer tissue obtained from previous studies^{11,12}. The most frequently altered genomic site identified was the *AR*-enhancer region, which was amplified in 81% of samples. The high prevalence of this type of amplification is notable because enhancer amplifications identified so far for other cancer types generally arise at much lower frequency^{13–16}. Moreover, the high prevalence of this *AR*-enhancer amplification in the data presented by Viswanathan and Quigley contrasts with its occurrence in only 1 of 54 previously published whole-genome sequences of prostate-cancer samples obtained before clinical treatment had commenced¹⁷. However, given the relatively small number of these whole-genome sequences from before treatment, it remains to be determined whether amplification of the *AR*-enhancer region usually arises at the time when cancers become treatment resistant, or

whether this alteration is already present in a subset of tumour cells before the cancer stops responding to treatment.

Quigley *et al.* did not report any correlation between the presence or absence of the *AR*-enhancer amplification and whether the cancer had progressed to the stage at which the patient received a second type of anti-androgen-pathway treatment, such as enzalutamide, after the first line of therapy had failed. Viswanathan *et al.* present sequencing data from three patients for whom tumour samples were available from before and after second-line anti-androgen-pathway treatment with enzalutamide; these data reveal that the samples from after treatment had an amplification of *AR* and the *AR* enhancer. If additional data from patients indicate a connection between the amplification of the *AR* enhancer and the emergence of treatment resistance, perhaps this amplification could be monitored as a biomarker of disease progression. Viswanathan and colleagues demonstrated that such alterations could be tracked through analysis of tumour DNA that is shed into patients' bloodstreams.

These studies highlight three important aspects of the way in which genomic analysis could illuminate our understanding of how cancers develop resistance to therapy. First, studying tumour samples obtained from patients before and during treatment might be the best way to understand how treatment resistance develops. Second, analysing a series of patient samples — such as biopsies or tumour DNA isolated from blood samples — during the course of therapy might help to reveal whether crucial DNA alterations arise during treatment or were already present in a subset of tumour cells before treatment. Quigley and colleagues' work with a large number of patient samples only partially addresses this. Analysis of patients over time

might also help to determine when therapy needs to be altered to try to prevent the development of treatment-resistant disease. Third, the technologies available for detecting genomic changes are rapidly improving, and the sequencing approaches used in the current studies can detect complex DNA alterations that were particularly challenging to determine using earlier techniques.

The genomic revolution that started with the Human Genome Project is reaching the cusp of a wave of detailed genomic studies that investigate how cancer evolves during treatment. Such progress represents another step closer to an era of precision medicine for cancer therapy. ■

Kellie A. Cotter and Mark A. Rubin
are in the Department for BioMedical Research, University of Bern, CH-3008 Bern, Switzerland, and at Inselspital, Bern, Switzerland. **M.A.R.** is also at Weill Cornell Medicine, New York, USA.
e-mail: mark.rubin@dbmr.unibe.ch

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
3. Takeda, D. Y. *et al.* *Cell* **174**, 422–432 (2018).
4. Viswanathan, S. R. *et al.* *Cell* **174**, 433–447 (2018).
5. Quigley, D. A. *et al.* *Cell* **174**, 758–769 (2018).
6. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. *Nature Genet.* **46**, 1160–1165 (2014).
7. Taplin, M.-E. *et al.* *N. Engl. J. Med.* **332**, 1393–1398 (1995).
8. Visakorpi, T. *et al.* *Nature Genet.* **9**, 401–406 (1995).
9. Armenia, J. *et al.* *Nature Genet.* **50**, 645–651 (2018).
10. Kumar, A. *et al.* *Nature Med.* **22**, 369–378 (2016).
11. Aggarwal, R. *et al.* *Eur. Urol. Focus* **2**, 469–471 (2016).
12. Holmes, M. G. *et al.* *J. Vasc. Interv. Radiol.* **28**, 1073–1081 (2017).
13. Glodzik, D. *et al.* *Nature Genet.* **49**, 341–348 (2017).
14. Herranz, D. *et al.* *Nature Med.* **20**, 1130–1137 (2014).
15. Shi, J. *et al.* *Genes Dev.* **27**, 2648–2662 (2013).
16. Zhang, X. *et al.* *Nature Genet.* **48**, 176–182 (2016).
17. Baca, S. C. *et al.* *Cell* **153**, 666–677 (2013).

a series of patient samples — such as biopsies or tumour DNA isolated from blood samples — during the course of therapy might help to reveal whether crucial DNA alterations arise during treatment or were already present in a subset of tumour cells before treatment. Quigley and colleagues' work with a large number of patient samples only partially addresses this. Analysis of patients over time might also help to determine when therapy needs to be altered to try to prevent the development of treatment-resistant disease. Third, the technologies available for detecting genomic changes are rapidly improving, and the sequencing approaches used in the current studies can detect complex DNA alterations that were particularly challenging to determine using earlier techniques.

The genomic revolution that started with the Human Genome Project is reaching the cusp of a wave of detailed genomic studies that investigate how cancer evolves during treatment. Such progress represents another step closer to an era of precision medicine for cancer therapy. ■

Kellie A. Cotter and Mark A. Rubin are in the Department for BioMedical Research, University of Bern, CH-3008 Bern, Switzerland, and at Inselspital, Bern, Switzerland. M.A.R. is also at Weill Cornell Medicine, New York, USA.
e-mail: mark.rubin@dbmr.unibe.ch

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).

2. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
3. Takeda, D. Y. *et al.* *Cell* **174**, 422–432 (2018).
4. Viswanathan, S. R. *et al.* *Cell* **174**, 433–447 (2018).
5. Quigley, D. A. *et al.* *Cell* **174**, 758–769 (2018).
6. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. *Nature Genet.* **46**, 1160–1165 (2014).
7. Taplin, M.-E. *et al.* *N. Engl. J. Med.* **332**, 1393–1398 (1995).
8. Visakorpi, T. *et al.* *Nature Genet.* **9**, 401–406 (1995).
9. Armenia, J. *et al.* *Nature Genet.* **50**, 645–651 (2018).
10. Kumar, A. *et al.* *Nature Med.* **22**, 369–378 (2016).
11. Aggarwal, R. *et al.* *Eur. Urol. Focus* **2**, 469–471 (2016).
12. Holmes, M. G. *et al.* *J. Vasc. Interv. Radiol.* **28**, 1073–1081 (2017).
13. Glodzik, D. *et al.* *Nature Genet.* **49**, 341–348 (2017).
14. Herranz, D. *et al.* *Nature Med.* **20**, 1130–1137 (2014).
15. Shi, J. *et al.* *Genes Dev.* **27**, 2648–2662 (2013).
16. Zhang, X. *et al.* *Nature Genet.* **48**, 176–182 (2016).
17. Baca, S. C. *et al.* *Cell* **153**, 666–677 (2013).

PRIMATE BIOLOGY

Embryonic role for a longevity protein

Monkeys genetically engineered to lack the gene *SIRT6* die a few hours after birth, displaying severe growth defects. This finding reveals a previously unknown role for the *SIRT6* protein in primate development. [SEE LETTER P.661](#)

SHOSHANA NAIMAN & HAIM Y. COHEN

For decades, biologists using model organisms such as mice and fruit flies have faced concerns about the relevance of their findings to humans. Using a model that is more evolutionarily similar to humans, such as another primate, could potentially close this frustrating gap. On page 661, Zhang *et al.*¹ use CRISPR–Cas9 gene-editing techniques to generate macaque monkeys lacking the gene *SIRT6*. Strikingly, they show that the *SIRT6* protein has a role in embryonic development in macaques that was not previously uncovered in mice.

Mammalian *SIRT6* removes acetyl groups from histone proteins. DNA is packaged around histones in the nucleus, and this deacetylation condenses the packaged DNA, suppressing gene expression². In mice, *SIRT6* is known to be a longevity protein that regulates many factors that alter during ageing, including genome stability, inflammation and metabolism². Indeed, overexpression of *SIRT6* in male mice leads to health improvements and extends lifespan³, whereas *SIRT6*-deficient mice die a few weeks after birth, displaying features of premature ageing⁴.

It is unknown whether *SIRT6* is involved in longevity in humans. However, recent data show⁵ that an inactivating mutation in human *SIRT6* causes overexpression of embryonic stem-cell genes, which leads to abnormal

development and severe brain defects, resulting in embryonic death. These findings suggest a previously unappreciated role for *SIRT6* in embryonic development, which should be considered separately from its role in ageing.

Zhang and colleagues used CRISPR–Cas9 to create one male and three female macaque embryos that did not express *SIRT6*. The females died shortly after birth and the male died in the middle of gestation. The absence of *SIRT6* caused severe, whole-body developmental delays. Compared with wild-type newborns, the mutants showed lower bone density, lower levels of subcutaneous fat and immature intestines and skeletal muscle. The authors also found that *SIRT6*-deficient monkeys had smaller brains owing to delayed neuronal maturation and an increase in the number of immature neural progenitor cells. Overall, the *SIRT6*-mutant animals were born much smaller than controls and showed gene-expression and morphological profiles closer to those of a typical three-month-old fetus than a full-term animal born after six months of gestation (Fig. 1).

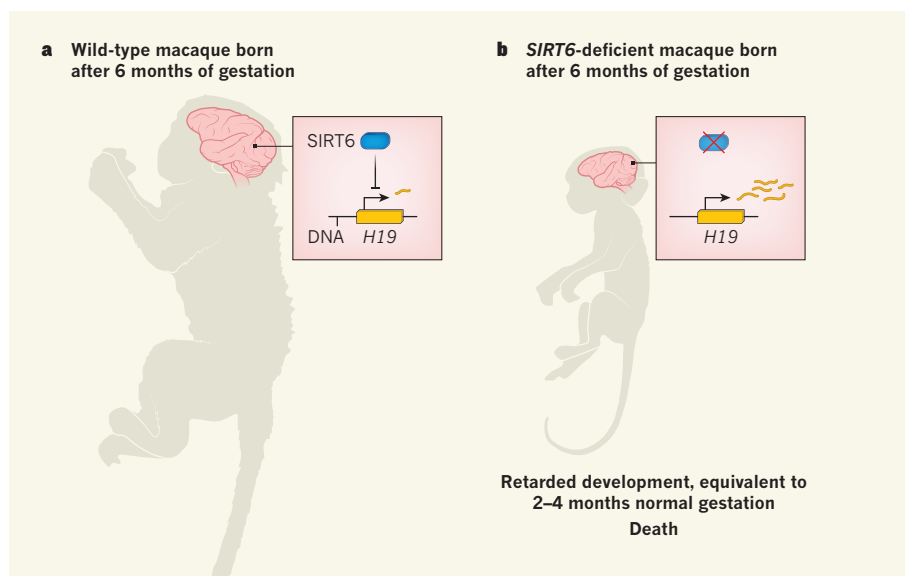


Figure 1 | A role for the *SIRT6* gene in primate development. Zhang *et al.*¹ used gene-editing tools to generate macaque embryos that did not express *SIRT6*. **a**, Wild-type macaques are born after around 6 months of gestation. In these animals, the *SIRT6* protein suppresses expression of the gene *H19*. **b**, *SIRT6*-deficient monkeys die a few hours after birth. These animals have major developmental defects and are a similar size to wild-type fetuses at 2 to 4 months of development. Notably, *SIRT6*-deficient animals have small and immature brains — this defect is accompanied by a dramatic increase in *H19* expression.

Because of the known role of SIRT6 in suppressing gene expression², Zhang *et al.* examined changes in gene expression in the mutants. Among the most upregulated genes was *H19*, which encodes a long non-coding RNA that is known to regulate fetal growth⁶. *H19* expression levels were increased in all tissues examined, with the highest expression in the brain.

Next, the authors used a different gene-editing approach to generate human neural progenitor cells lacking *SIRT6* *in vitro*, and showed that the differentiation of these cells into neurons was delayed when compared with wild-type cells. This defect was accompanied by higher levels of *H19* RNA. Finally, the group found that SIRT6 removes acetyl groups associated with *H19* transcription, and showed that reducing *H19* expression in human cells lacking *SIRT6* resolved their defects in neuronal differentiation. Thus, SIRT6 inhibits *H19* expression to modulate neuronal development in human cells, as in monkeys.

Several avenues for further work arise from these results. For instance, the absence of SIRT6 altered the expression of thousands of genes in various tissues, and it is unlikely that *H19* is the only gene responsible for the defects observed. Indeed, a human developmental disorder called Silver–Russell syndrome can be caused by increased *H19* levels but, in contrast to *SIRT6*-deficient monkeys, people who have this disorder have normal lifespans and less-severe developmental changes⁶. This discrepancy suggests that SIRT6-modulated genes other than *H19* also contribute to the severe effects seen in the authors' mutant monkeys. It will be hard to pinpoint the precise genes that cause developmental defects in *SIRT6*-deficient animals, but this should be investigated in the future.

From an evolutionary point of view, *SIRT6* is fascinating. In all mammals studied, the gene's deletion causes premature death, and the protein has the same enzymatic activity and involvement in glucose metabolism and stem-cell differentiation⁷. However, as we climb the evolutionary ladder from mice to monkeys to humans, some of the traits caused by *SIRT6* deletion become progressively more severe. *SIRT6*-deficient mice die a few weeks to months after birth⁸, whereas monkeys die within hours, and humans harbouring a *SIRT6*-inactivating mutation are not even born. This increasing severity could be explained by the acquisition of regulatory roles for *SIRT6* over the course of evolution. In support of this idea, the severe brain defects seen in *SIRT6*-deficient primates have not been reported in mice, and this change correlates well with differences in brain complexity in these species. It will be extremely interesting to further explore the source of this trait enhancement across evolution.

What can we learn about the role of SIRT6 in human ageing from this primate model? At first glance, there is not an obvious connection

between the developmental defects seen in the monkeys and ageing, as they are at opposite ends of life's timeline. However, key pathways regulated by SIRT6 are conserved between these species, and genome-wide association studies have found a correlation between *SIRT6* and increased lifespan in humans⁹. These facts, together with data indicating that SIRT6 helps to protect the brain against ageing-related disorders such as Alzheimer's disease¹⁰, strongly suggest that the versatile SIRT6 protein might promote healthy longevity in humans. In the future, developments in CRISPR engineering might enable gene editing in specific tissues, and at chosen time points; if the latter were achieved, it would be fascinating to characterize the role of SIRT6 in primate lifespan.

More generally, genome editing is an exciting future strategy for human therapy. However, the challenge is to induce the desired edits without creating nonspecific mutations or producing mosaic embryos in which only some cells express the edited gene. Promisingly, Zhang and colleagues found no mosaicism or detectable off-target mutations in their mutant animals, and another group that have used

CRISPR in monkeys also report no off-target effects¹¹. Although there are still many ethical and technical caveats to be considered, the authors' achievement — along with a similar success in human embryos¹² — gives hope that human genetic therapies using CRISPR engineering will be possible in the future. ■

Shoshana Naiman and Haim Y. Cohen are at the Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel.

e-mail: haim.cohen@biu.ac.il

1. Zhang, W. *et al.* *Nature* **560**, 661–665 (2018).
2. Tennent, R. I. & Chua, K. F. *Trends Biochem. Sci.* **36**, 39–46 (2011).
3. Kanfi, Y. *et al.* *Nature* **483**, 218–221 (2012).
4. Mostoslavsky, R. *et al.* *Cell* **124**, 315–329 (2006).
5. Ferrer, C. M. *et al.* *Genes Dev.* **32**, 373–388 (2018).
6. Abu-Amero, S. *et al.* *J. Med. Genet.* **45**, 193–199 (2008).
7. Tasselli, L., Zheng, W. & Chua, K. F. *Trends Endocrinol. Metab.* **28**, 168–185 (2017).
8. Peshti, V. *et al.* *PLoS One* **12**, e0176371 (2017).
9. Hirvonen, K. *et al.* *BMC Med. Genet.* **18**, 41 (2017).
10. Kaluski, S. *et al.* *Cell Rep.* **18**, 3052–3062 (2017).
11. Zuo, E. *et al.* *Cell Res.* **27**, 933–945 (2017).
12. Ma, H. *et al.* *Nature* **548**, 413–419 (2017).

This article was published online on 22 August 2018.

STRUCTURAL BIOLOGY

Transcriptional speed bumps revealed

The enzyme RNA polymerase II, which transcribes DNA, pauses early in transcription and awaits signals to continue. High-resolution structures reveal how it is stopped and efficiently restarted. SEE ARTICLES P.601 & P.607

KAREN ADELMAN & TELMO HENRIQUES

A first step in gene expression is the recruitment of the DNA-transcribing enzyme RNA polymerase II (Pol II) to a gene, and the assembly of transcriptional machinery around it. Pol II can then initiate RNA synthesis. However, during transcription of most mammalian genes, Pol II does something peculiar — after synthesizing a short RNA molecule usually no longer than 60 nucleotides, it stops, awaiting further instructions before transcribing the remainder of the gene¹. Such pausing and subsequent RNA elongation is central to gene regulation in animals, yet the mechanisms underlying this process have not been clear. In two papers in this issue, Vos *et al.*^{2,3} describe landmark structures that shed new light on Pol II pausing and release.

A heterodimer comprising the proteins SPT4 and SPT5 is crucial for the pausing of Pol II (ref. 4). During transcription initiation, general transcription factors bind and occlude the regions of Pol II recognized

by SPT5 — these factors must be released before SPT5 can associate. Thus, SPT5 binding occurs after transcription proper begins, and stable interactions between SPT5 and Pol II require a nascent RNA about 20 nucleotides in length to have formed⁵. Interactions with transcribing Pol II then enable SPT5 to recruit additional factors that govern Pol II activity and RNA processing^{4,5}. One such factor is the negative elongation factor (NELF) protein complex, which comprises four subunits (NELF-A, -B, -C and -E)⁴.

In contrast to SPT5, which is evolutionarily conserved from bacteria all the way through to humans, no equivalents to the mammalian NELF proteins have been identified in bacteria, yeast, worms or plants⁴. The organisms that do contain a NELF complex are those that exhibit stable pausing of Pol II, implying a role for NELF in this process. Indeed, the release of NELF from Pol II is concomitant with escape from pausing into elongation¹, and acute depletion of NELF both prevents normal pausing⁶ and increases premature termination⁷.

(the process whereby Pol II inadvertently releases DNA, ceasing transcription). But the molecular basis of NELF activity has remained obscure. In particular, it has been unclear how NELF interacts with Pol II and how it might stabilize the paused state in a manner that prevents both continued RNA synthesis and transcription termination.

In the first of their papers (page 601), Vos *et al.*² used cryo-electron microscopy to resolve the structure of a paused transcription complex at 3.2-ångström resolution. The authors assembled a highly purified structure on an artificial DNA–RNA scaffold that contains sequences known⁸ to strongly promote Pol II pausing, using pig Pol II along with human SPT5 and NELF complexes. The Pol II–SPT5–NELF complexes formed on this scaffold showed clear differences compared with previously published Pol II–SPT5 complexes in an actively transcribing conformation⁹. Whereas the DNA–RNA hybrid held within active Pol II has an unpaired DNA base that can be used as a template to direct addition of the next RNA nucleotide, the DNA–RNA hybrid in the paused complex is ‘tilted’ and lacks unpaired template DNA. Without a free DNA base in its active site, Pol II is unable to carry out RNA elongation.

This non-productive DNA–RNA hybrid conformation alone explains why Pol II pauses. But more importantly, the structure also reveals the role of NELF in this process. The researchers found that a protein lobe comprising the NELF-A and NELF-C subunits binds near a funnel region in Pol II through which nucleotides normally access the active site. The NELF lobe protrudes into the funnel, potentially restricting the entry of nucleotides needed for transcription. In addition, NELF restrains mobile loop domains in Pol II, such as the trigger loop, near the active site. This restraint locks the enzyme in the inactive conformation while simultaneously discouraging Pol II from sliding along the DNA, which can lead to transcription termination.

The NELF binding pocket near the Pol II funnel overlaps with a region that, when not occluded, can be bound by the factor TFIIS to stimulate elongation. Intriguingly, TFIIS has been shown to reactivate Pol II that adopts a non-productive, tilted DNA–RNA hybrid conformation¹⁰. Thus, Vos *et al.* propose that NELF also prevents Pol II reactivation by blocking TFIIS binding (Fig. 1).

The release of paused Pol II into elongation is triggered by the recruitment of the kinase enzyme P-TEFb, which phosphorylates Pol II and pause-inducing factors, triggering dissociation of NELF (ref. 1). P-TEFb activity is accompanied by the recruitment to Pol II of the SPT6 protein and the polymerase-associated factor (PAF) protein complex. However, whether these elongation-associated factors directly affect Pol II pause release has been unclear. In the second of the papers (page 607), Vos *et al.*³ examined this possibility

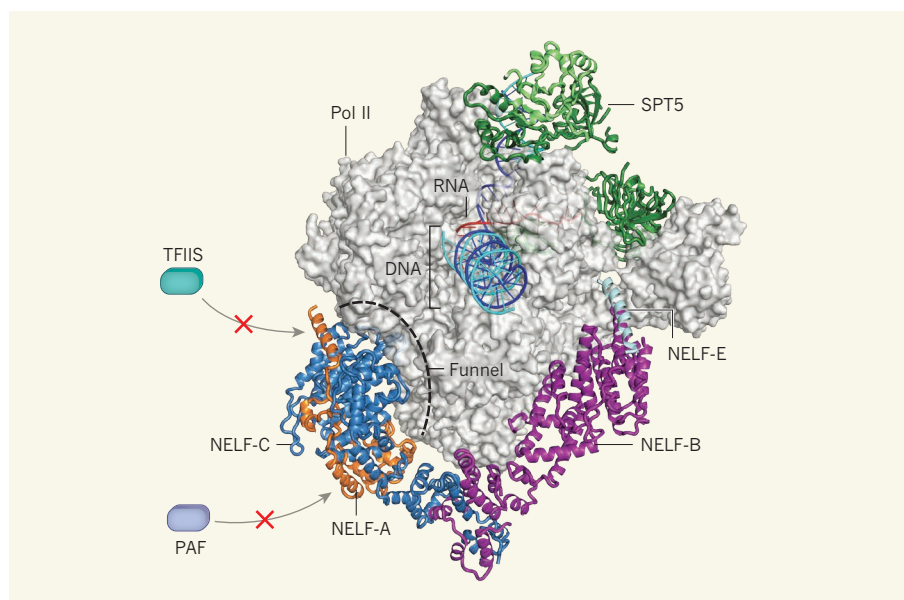


Figure 1 | Visualizing transcriptional pause and release. The DNA-transcribing enzyme RNA polymerase II (Pol II) pauses after initiating RNA synthesis and must be reactivated to continue transcriptional elongation of the nascent RNA. Vos *et al.*^{2,3} solved high-resolution structures of Pol II and the transcriptional elongation machinery around it, both in the paused state and after elongation has resumed (the latter is not shown here). In the paused state, nascent RNA and the DNA being transcribed are held by the SPT5 protein and two subunits of the NELF protein complex (NELF-A and NELF-C) in a tilted conformation that prevents further transcription. NELF binds close to a funnel domain in Pol II, and blocks binding of Pol II by the factors TFIIS and PAF, which are needed for efficient elongation. NELF dissociates from Pol II to allow this binding to occur in the reactivated complex. (Figure adapted from Fig. 2b of ref. 2.)

by assembling a structure that included a modified, elongation-permissive nucleic-acid scaffold and these activating proteins.

As anticipated, the DNA–RNA hybrid in the activated elongation complex is no longer tilted and adopts a conformation compatible with RNA synthesis. The authors found multiple sites phosphorylated by P-TEFb in both SPT5 and NELF. Phosphorylation at these sites might aid the opening of the interface between Pol II and SPT5, and lead to dissociation of NELF. Furthermore, the group showed that phosphorylation of SPT6 and a linker region in the carboxy-terminal domain of Pol II aided docking of SPT6 on the enzyme. Most strikingly, the structure revealed that the binding of NELF and PAF to Pol II is mutually exclusive. Thus, dissociation of NELF during pause release enables the binding of PAF as well as TFIIS, allowing transcription to proceed.

Taking these results together, a detailed molecular model of Pol II pausing and release begins to emerge. We note a recurring theme wherein mutually exclusive, overlapping binding sites for a succession of Pol II-associated factors enable an orderly exchange during the transcription cycle. Furthermore, the specificity of each protein's interaction with the Pol II complex is ensured by multiple interaction interfaces, often with scaffold proteins such as SPT5 and the nucleic acids.

Of course, questions remain about the transition from pausing to productive elongation. For example, this work calls into question

the roles of RNA-binding domains found in NELF subunits⁴. Surprisingly, Vos *et al.* showed that disruption of one such domain in NELF-E had no effect on pausing. It also remains to be seen whether the tilted DNA–RNA conformation observed by the authors is prevalent *in vivo*, and how the phosphorylation of pause-inducing factors drives pause release.

This work represents a fundamental jump in our understanding of pausing. The structures point to several appealing models for regulated pause release that can be tested in future work. ■

Karen Adelman and Telmo Henriques are in the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. e-mail: karen_adelman@hms.harvard.edu

- Adelman, K. & Lis, J. T. *Nature Rev. Genet.* **13**, 720–731 (2012).
- Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. *Nature* **560**, 601–606 (2018).
- Vos, S. M. *et al.* *Nature* **560**, 607–612 (2018).
- Yamaguchi, Y., Shibata, H. & Handa, H. *Biochim. Biophys. Acta* **1829**, 98–104 (2013).
- Missra, A. & Gilmour, D. S. *Proc. Natl Acad. Sci. USA* **107**, 11301–11306 (2010).
- Core, L. J. *et al.* *Cell Rep.* **2**, 1025–1035 (2012).
- Henriques, T. *et al.* *Mol. Cell* **52**, 517–528 (2013).
- Palangat, M., Meier, T. I., Keene, R. G. & Landick, R. *Mol. Cell* **1**, 1033–1042 (1998).
- Bernecky, C., Plitzko, J. M. & Cramer, P. *Nature Struct. Mol. Biol.* **24**, 809–815 (2017).
- Cheung, A. C. M. & Cramer, P. *Nature* **471**, 249–253 (2011).

This article was published online on 22 August 2018.

FUNDAMENTAL CONSTANTS

Gravity measured with record precision

The gravitational constant, G , which governs the strength of gravitational interactions, is hard to measure accurately. Two independent determinations of G have been made that have the smallest uncertainties so far. [SEE ARTICLE P.582](#)

STEPHAN SCHLAMMINGER

Although gravity seems strong in our everyday lives, such as when lifting a heavy object, it is the weakest of the four fundamental forces. The gravitational force between two bodies is proportional to the masses of these bodies. If one of the bodies is Earth, the force can be considerable. But if the bodies are objects in a laboratory, the force can be too small to measure accurately. For example, the gravitational force between two 1-kilogram objects separated by 1 metre is equivalent to the weight of a few biological cells. For this reason, the gravitational constant, G , which quantifies the strength of this force, is one of the most poorly defined physical constants. But on page 582, Li *et al.*¹ report high-precision measurements of G using two different techniques.

In 1798, the scientist Henry Cavendish determined G for the first time in the

laboratory, using an instrument called a torsion balance². In Cavendish's work, the torsion balance consisted of a dumb-bell that was suspended from its centre by a thin fibre. A gravitational force was applied to the masses at the ends of the dumb-bell, acting perpendicularly to the bar of the dumb-bell and to the axis of the fibre. This force led to a rotation of the dumb-bell about this axis, causing the fibre to twist.

Eventually, the dumb-bell reached a position at which the twisting force of the fibre balanced the gravitational force. The rotation angle of the dumb-bell in this position was recorded. The gravitational force was then applied in the opposite direction and a second rotation angle was measured. The magnitude of the gravitational force was calculated from the difference between these two angles.

In torsion-balance experiments, the gravitational force is provided by a well-characterized assembly of external masses. These masses are

moved between two or more different positions to change the direction and magnitude of the force. Because the dumb-bell rotates in a horizontal plane, the otherwise overwhelming effects of Earth's gravity on the experiments are negligible. Over the years, many techniques have been developed to measure G using a torsion balance³. In 2000, a substantial improvement in the precision of these experiments was achieved by replacing the dumb-bell with a thin plate⁴ (also termed a plate pendulum).

Li and colleagues built two plate-containing torsion balances that are based on different measurement techniques: the time-of-swing (TOS) method⁵ and the angular-acceleration-feedback (AAF) method⁶ (see Fig. 1 of the paper¹). In the TOS method, the rotation of the plate is oscillatory. G is calculated from the change in the speed of the oscillation when the external masses are in two different configurations. By contrast, in the AAF method, two turntables are used to rotate the torsion balance and the external masses individually. G is determined from the angular acceleration of the turntable associated with the torsion balance when the amount of twisting of the fibre is reduced to zero.

The authors obtained G values of 6.674184×10^{-11} and 6.674484×10^{-11} cubic metres per kilogram per square second for the TOS method and the AAF method, respectively. The relative uncertainties are the smallest reported so far: about 11.6 parts per million. By comparison, the previous record, which was achieved using the AAF method, was 13.7 parts per million⁴.

Li *et al.* carried out their experiments with great care and gave a detailed description of their work. The study is an example of excellent craftsmanship in precision measurements. However, the true value of G remains unclear. Various determinations of G that have been made over the past 40 years have a wide spread of values (Fig. 1). Although some of the individual relative uncertainties are of the order of 10 parts per million, the difference between the smallest and largest values is about 500 parts per million.

There are at least two possible explanations for this discrepancy. One is that the technical details of one or more of the experiments were not fully understood, which could have led either to a systematic shift in the reported values of G or to uncertainties that were not included in the reported uncertainties of G . An example of the former is the effect of a fibre property, called anelasticity, that could bias the TOS method — an effect that was first pointed out⁷ in 1995. A second possibility is that some unknown physics could explain the scatter in the published values. Although this possibility is, of course, the more exciting, it is also the less likely. Nevertheless, it should not be dismissed lightly.

At this point, it is as important to try to understand the discrepancy between the different results as it is to make new

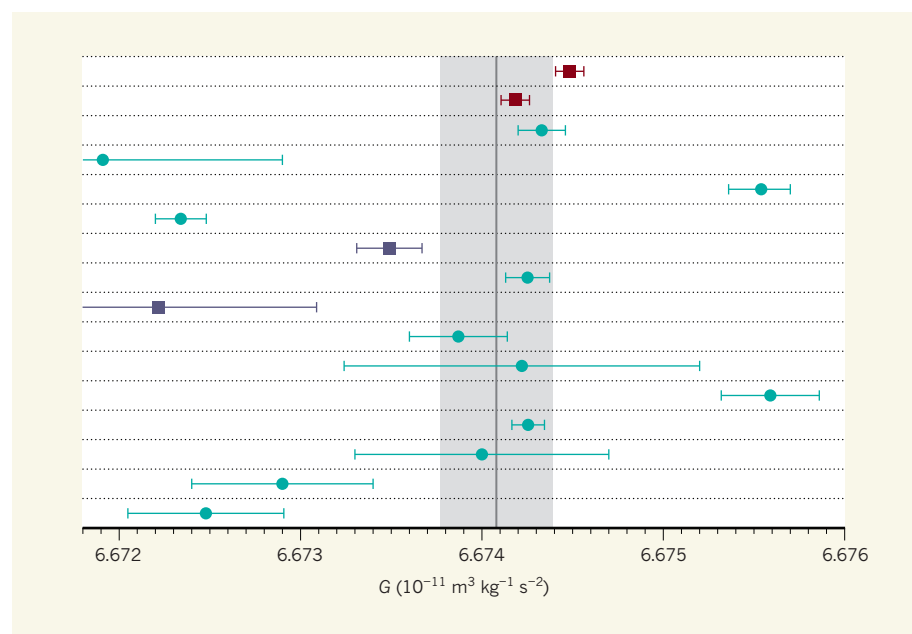


Figure 1 | Measurements of the gravitational constant. The strength of the gravitational force between two bodies is described by the gravitational constant, G , which can be expressed in units of cubic metres per kilogram per square second. The data points are high-precision measurements of G taken over the past 40 years, with uncertainties indicated by the error bars. The points marked by squares are results obtained by Li *et al.* in current work¹ (red) and in previous work^{8,9} (purple). The vertical grey line denotes the value of G adopted by the Committee on Data for Science and Technology, with an uncertainty indicated by the shaded area¹¹. (Adapted from Fig. 3 of ref. 1.)

measurements. Even Li and colleagues' results are in disagreement: the values of G determined in the two current experiments, as well as values obtained in two previous experiments at the same laboratory^{8,9}, are statistically inconsistent with one another. The authors speculate that fibre anelasticity might be responsible, but they do not give a definitive explanation.

Because all four of these experiments were carried out at the same institution, it should be more straightforward to compare them than it would be to compare different experiments from various groups around the globe. An excellent opportunity exists, therefore, to uncover the causes of the discrepancy and, in turn, to learn more about the true value of G . Li *et al.* should be encouraged to take on this challenge. In the end, if we want to understand the measurements of G , we must find the reasons for the inconsistent results¹⁰. ■

Stephan Schlamminger is in the Fundamental Electrical Measurements Group, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA. e-mail: stephan.schlamminger@nist.gov

1. Li, Q. *et al.* *Nature* **560**, 582–588 (2018).
2. Cavendish, H. *Phil. Trans. R. Soc. B* **88**, 469–526 (1798).
3. Rothleitner, C. & Schlamminger, S. *Rev. Sci. Instrum.* **88**, 111101 (2017).
4. Gundlach, J. H. & Merkowitz, S. M. *Phys. Rev. Lett.* **85**, 2869–2872 (2000).
5. Reich, F. *Abh. Math.-Phys. Cl. Königliche Sächsischen Ges. Wiss.* **1**, 384–430 (1852).
6. Rose, R. D., Parker, H. M., Lowry, R. A., Kuhlthau, A. R. & Beams, J. W. *Phys. Rev. Lett.* **23**, 655–658 (1969).
7. Kuroda, K. *Phys. Rev. Lett.* **75**, 2796 (1995).
8. Hu, Z. K., Guo, J. Q. & Luo, J. *Phys. Rev. D* **71**, 127505 (2005).
9. Tu, L. C. *et al.* *Phys. Rev. D* **82**, 022001 (2010).
10. Quinn, T. *Nature* **505**, 455 (2014).
11. Mohr, P. J., Taylor, B. N. & Newell, D. B. *Rev. Mod. Phys.* **88**, 035009 (2016).

PLANT GENETICS

A new green revolution on the horizon

Manipulation of the transcription factor OsGRF4 can improve the efficiency with which some high-yielding cereal crops use nitrogen. This discovery has implications for sustainable agriculture. [SEE ARTICLE P.595](#)

FANMIAO WANG & MAKOTO MATSUOKA

The green revolution of the mid-twentieth century saw the development of high-yielding varieties of rice and wheat for use in agriculture. But to produce high yields, these green-revolution varieties require a large supply of nitrogen. Developing green-revolution varieties that use nitrogen more efficiently is an important goal for sustainable crop breeding. On page 595, Li *et al.*¹ report a previously unknown function for the rice transcription factor OsGRF4 in nitrogen use. By modulating the *OsGRF4* gene, the researchers produced plants that use nitrogen efficiently and have a high yield.

Proteins of the DELLA family inhibit plant growth, whereas hormones called gibberellins promote plant growth by triggering the destruction of DELLA proteins. Green-revolution varieties of rice and wheat harbour genetic mutations that lead to the accumulation of DELLA proteins. As a result, these plants are shorter than are normal varieties, and so are resistant to lodging^{2,3} — the process by which plants are flattened by wind and rain. This lodging resistance is a fundamental mechanism for achieving increased crop yield in green-revolution varieties.

DELLA accumulation also inhibits nitrogen uptake and nitrogen-related growth

responses — traits that are associated with the inefficient use of nitrogen⁴. Consequently, farmers have to apply large amounts of environmentally damaging nitrogen-based fertilizer to their crops to achieve high yields in green-revolution varieties. Although DELLA accumulation increases the yield, it therefore also has a negative impact in terms of sustainable agriculture.

Li *et al.* set out to overcome the negative impact of DELLA accumulation. They crossed varieties of the rice subspecies *Oryza sativa indica* that showed differing rates of nitrogen uptake. They then performed genetic analyses on the resulting plants, which had a range of yields. In doing so, they found that *OsGRF4* is associated with nitrogen uptake. *OsGRF4* has previously been found to regulate the size of rice grains^{5–7} and the levels of growth molecules called cytokinins⁸, both of which affect crop yield. But no relationship between *OsGRF4* and nitrogen-use efficiency has previously been described.

The researchers genetically engineered green-revolution varieties of rice to lack *OsGRF4*. Compared with control plants carrying the wild-type gene, mutants showed less nitrogen-dependent growth and reduced nitrogen uptake and assimilation (the process by which inorganic nitrogen from fertilizers is converted into useful organic compounds such



50 Years Ago

Mr J. H. Brazell of the Meteorological Office has compiled a book of weather statistics for the London area which promises to become a well-thumbed reference ... The year 1841 ... is the first year for which regular official meteorological observations are available ... Mr Brazell has taken this opportunity to delve into earlier chronicles to find what London's weather was like before 1841 ... A rare feature of London's climate has been the freezing of the Thames ... During twenty-three winters between 1260 and 1814, the ice on the river was thick enough to allow pedestrians to cross from one bank to the other. It became the custom for frost fairs to be held on the frozen Thames, starting from small beginnings in the winter of 1309–10 when people danced around a bonfire built on the ice, to the great frost fairs of the 17th, 18th and 19th centuries, when the frozen river supported streets of shops and booths.

From *Nature* 31 August 1968

100 Years Ago

The July issue of *Science Progress* contains an interesting article by Sir Henry Thompson on the food requirements of a normal working-class family. A comparison is instituted between the physiological values of the diets reported upon by the Board of Trade in pre-war times and some data collected by the War Emergency Committee in 1917 ... Sir Henry Thompson has employed a more liberal scale of requirements for children than the older standard of Atwater, which is now generally recognised to be unsatisfactory. The three diets do not differ greatly in respect of energy-value; the highest average is that of the urban working-class families (1913), yielding 3410 calories; the lowest, the 1917 sample, is 3160 calories, a reduction of but 250 calories.

From *Nature* 29 August 1918

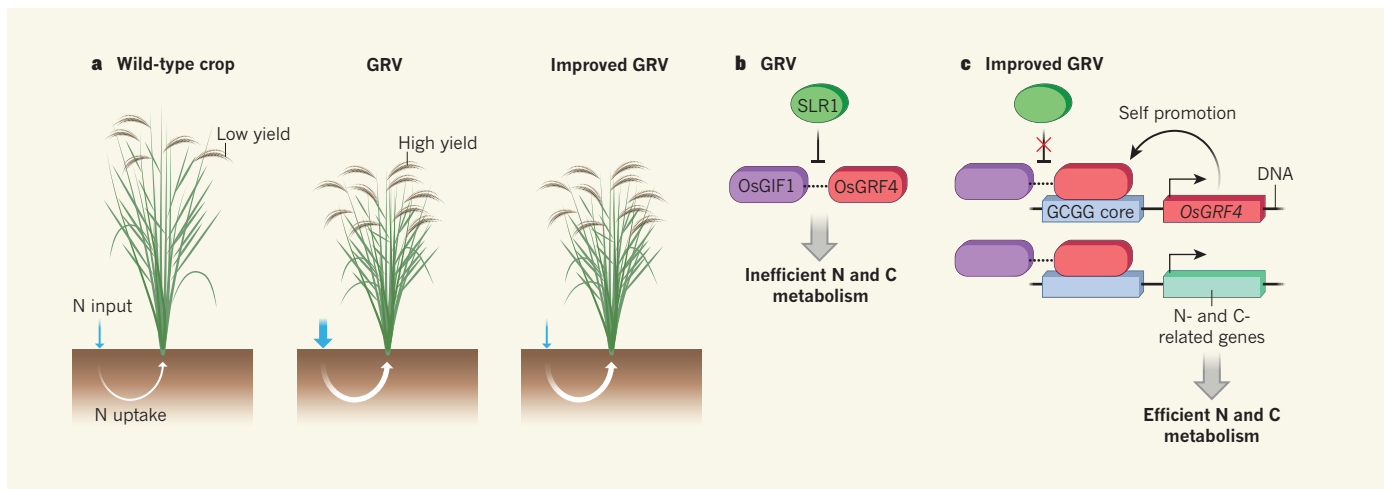


Figure 1 | Tipping the scales to improve plant yields. **a**, Nitrogen is a major component of fertilizers. The height of many wild-type crop plants makes them prone to flattening by wind and rain if they are grown under high nitrogen (N) input. Therefore, farmers cultivate wild-type crops under low nitrogen input, which decreases yield. Yields can be improved by generating short plants known as green-revolution varieties (GRVs), in which the growth-inhibiting protein SLR1 accumulates (not shown). However, GRVs take up and use nitrogen inefficiently, and so require high levels of nitrogen input to take up sufficient nitrogen to produce high yields. Li *et al.*¹ have generated improved GRVs that both have high crop yields and

use nitrogen efficiently. **b**, The authors found that, in GRVs, SLR1 inhibits the interaction between two proteins, OsGIF1 and OsGRF4. This reduces the efficiency with which both nitrogen and carbon (C) are metabolized. **c**, The group bred plants that produced high levels of OsGRF4, thereby overcoming the ability of SLR1 to prevent OsGIF1–OsGRF4 interactions. In these improved GRVs, OsGRF4 binds to a specific DNA sequence (the GCGG core) to promote the expression of the *OsGRF4* gene, and of genes involved in nitrogen and carbon use. This leads to a feed-forward loop that increases the efficiency of nitrogen and carbon metabolism and results in higher yields.

as amino acids). By contrast, plants that were selectively bred to express *OsGRF4* at higher than normal levels showed an increased rate of nitrogen uptake. Thus, *OsGRF4* promotes various nitrogen-related events.

Li *et al.* then demonstrated that *OsGRF4* acts in opposition to the DELLA protein SLR1 in rice (Fig. 1). Transcriptional activation by *OsGRF4* is known to be promoted by physical interactions between *OsGRF4* and another protein, *OsGIF1* (refs 5 and 6). The authors found that *OsGRF4*, promoted by *OsGIF1*, binds to a specific DNA sequence (the core motif GCGG) to drive the expression of genes that encode a range of proteins involved in nitrogen metabolism, uptake and assimilation. However, the accumulation of SLR1, as occurs in green-revolution varieties of rice, inhibits the interaction between *OsGRF4* and *OsGIF1*, thereby suppressing the expression of the genes involved in nitrogen uptake and metabolism. This SLR1-mediated inhibition was relieved by the presence of gibberellin. Li and colleagues also showed that the expression of *OsGRF4* itself is activated by the *OsGRF4*–*OsGIF1* complex. Therefore, *OsGRF4* transcription is suppressed by SLR1.

Next, the group discovered that *OsGRF4* and SLR1 have the same antagonistic relationship in another key process in plant metabolism, carbon assimilation. Products of carbon and nitrogen assimilation act together to form the building blocks needed for metabolic processes in plants, and a balance between the two is therefore essential for optimal growth and yield. The authors showed that *OsGRF4* promotes, and that SLR1 inhibits, the expression of various genes that are

involved in three carbon-related processes: photosynthesis, sucrose transport and sucrose metabolism. Furthermore, the same relationship governs the expression of several genes involved in cell-cycle progression. Li and colleagues therefore propose that the antagonistic relationship between *OsGRF4* and SLR1 provides a regulatory link that coordinates plant growth, nitrogen metabolism and carbon assimilation.

Finally, the authors used their findings to improve the yields of green-revolution varieties. They applied breeding strategies to generate rice plants that produce high levels of *OsGRF4* but retain the short stature of green-revolution varieties. The resulting plants had wider leaves and stems, and showed an increased nitrogen uptake compared with normal plants. Consequently, crop yield was increased, even at low levels of nitrogen; an optimal carbon–nitrogen balance was attained; and the plants maintained their beneficial stature. Li *et al.* achieved a similar effect by increasing *OsGRF4* expression in the rice subspecies *Oryza sativa japonica*, as well as in green-revolution varieties of wheat. They have therefore successfully disconnected gibberellin-mediated control of plant height from regulation of nitrogen metabolism, producing plants that grow better without an increased risk of lodging.

Li and colleagues' study raises the question of how increased levels of *OsGRF4* can increase

plant growth horizontally (by increasing leaf and stem width) but not vertically (through stem elongation). Answering this question will involve in-depth studies of nitrogen-related genes that are under the regulatory control of *OsGRF4*.

More importantly, the authors' work not only reminds us of the disadvantages of green-revolution varieties, but also demonstrates that they can be overcome by implementing breeding strategies to increase levels of *OsGRF4*. By improving the efficiency of nitrogen use by green-revolution varieties, the amount of nitrogen-based fertilizers that are needed for agriculture could be reduced, which would improve our ability to grow crops sustainably. Li and colleagues' research should also stimulate the discovery of other genes and molecules with roles in nitrogen use that are independent of gibberellin-regulated plant growth. Identifying fresh targets for breeding strategies in this way could usher in a new green revolution. ■

Fanmiao Wang and Makoto Matsuoka are at the Bioscience and Biotechnology Center, Nagoya University, Nagoya 464-8601, Japan. e-mail: makoto@nuagr1.agr.nagoya-u.ac.jp

- Li, S. *et al.* *Nature* **560**, 595–600 (2018).
- Peng, J. *et al.* *Nature* **400**, 256–261 (1999).
- Sasaki, A. *et al.* *Nature* **416**, 701–702 (2002).
- Gooding, M. J., Addisu, M., Uppal, R. K., Snape, J. W. & Jones, H. E. *J. Agric. Sci.* **150**, 3–22 (2012).
- Che, R. *et al.* *Nature Plants* **2**, 15195 (2015).
- Duan, P. *et al.* *Nature Plants* **2**, 15203 (2015).
- Hu, J. *et al.* *Mol. Plant* **8**, 1455–1465 (2015).
- Sun, P. *et al.* *J. Integr. Plant Biol.* **58**, 836–847 (2016).

This article was published online on 15 August 2018.

Subwavelength integrated photonics

Pavel Cheben^{1*}, Robert Halir^{2,3}, Jens H. Schmid¹, Harry A. Atwater⁴ & David R. Smith⁵

In the late nineteenth century, Heinrich Hertz demonstrated that the electromagnetic properties of materials are intimately related to their structure at the subwavelength scale by using wire grids with centimetre spacing to manipulate metre-long radio waves. More recently, the availability of nanometre-scale fabrication techniques has inspired scientists to investigate subwavelength-structured metamaterials with engineered optical properties at much shorter wavelengths, in the infrared and visible regions of the spectrum. Here we review how optical metamaterials are expected to enhance the performance of the next generation of integrated photonic devices, and explore some of the challenges encountered in the transition from concept demonstration to viable technology.

A periodic crystal lattice acts like a diffraction grating for X-rays with wavelengths comparable to the lattice constant, but appears like a homogeneous medium for light of the much longer optical wavelengths. Similarly, a dielectric grating can diffract light or behave as an equivalent homogeneous medium, depending on the ratio of the wavelength of the light to the periodicity of the grating. In a subwavelength grating (SWG), the fundamental dielectric building blocks, which are arranged periodically, assume the role of the atoms of the crystal lattice and ultimately determine the macroscopic optical properties of the metamaterial. Indeed, if the period of the grating is much smaller than the wavelength of the light, diffraction effects are suppressed, and the structure behaves like a homogeneous anisotropic material with an equivalent anisotropic permittivity tensor¹ with respect to the macroscopic electromagnetic field. Artificial media with optical properties synthesized by deliberate structuring have been used for over 50 years in diffractive free-space optics^{2,3}. Some early subwavelength structures were also used in semiconductor multilayers⁴ and waveguides⁵ for phase-matched nonlinear frequency conversion. The term ‘metamaterial’ was coined more recently^{6–8}, and originally referred to artificial media designed to have a greater range of material properties than those available in nature. Metamaterials based on metallic structures were subsequently developed to demonstrate exotic properties—such as negative permeability and permittivity⁹, super-resolution⁷, invisibility¹⁰ and asymmetric transmission¹¹—or in the quest for optical magnetism¹². Current metamaterial research includes the study of metallic, hybrid metallic–dielectric and all-dielectric nanostructures, leading to new photonic device concepts, which have been described in several comprehensive review articles^{13–22}.

In this review we discuss how bringing metamaterials into optical-waveguide technologies and on-chip architectures provides new degrees of freedom to control the flow of light in integrated photonic devices. We emphasize the role of SWGs in silicon-based integrated optical circuits²³, which are considered to be key components for the development of the next generation of optical communication, biomedical, quantum and sensing technologies.

Subwavelength-grating metamaterial structures were recently implemented in silicon waveguides^{24–26}, allowing accurate lithographic control over the distribution of the electromagnetic field and the wavevector of the propagating modes²⁷. Through the realization of practical components at telecommunication wavelengths, it was demonstrated that waveguide mode transformation can be controlled by changing the effective material index, achieving a broad wavelength range with a negligible level of scattering loss^{28,29}. Independently, Levy et al.³⁰ showed that a spatially

inhomogeneous metamaterial can be used to control the effective index of refraction in a silicon slab waveguide. A unique aspect of the slab waveguide configuration is the large degree of control in creating a wide range of different spatial distributions of metamaterial refractive index by lithographic nano-patterning. This level of control has been demonstrated on various integrated structures, including the waveguide lens³⁰, the invisibility cloak³¹, a flattened Luneburg lens³², Maxwell’s fish-eye lens³³ and dual-function ‘Janus’ devices³⁴.

The emerging opportunity to control the properties of integrated optical structures at the subwavelength scale has motivated intense research efforts, and a plethora of advanced devices with unprecedented performance have been demonstrated^{27,28,30,35–42}. Such subwavelength devices can be fabricated in the same lithography step as conventional waveguides by using manufacturing processes that are well established in the semiconductor electronics industry, thus making their integration straightforward. Highly efficient subwavelength structures for coupling light into integrated photonic devices have been developed, including subwavelength-engineered edge couplers^{36,43} and surface grating couplers hybridized with optical metasurfaces^{40,44,45} at both near-infrared (telecommunication) and mid-infrared wavelengths. Subwavelength systems for sensing^{39,46,47}, and even an electronic–photonic system integrating transistors and nanostructured optical elements⁴⁸ on a single chip, have been demonstrated.

In the following, we review diverse implementations of subwavelength-engineered structures in integrated optics. We begin by summarizing the physical principles of SWG metamaterial structures related to the operation of integrated photonic platforms. Next, we describe the state of the art of metamaterial devices in silicon-on-insulator waveguides and analyse the arising challenges vis-à-vis the development of viable photonic integrated technology. We emphasize the need for functional metamaterial photonic elements that can be integrated on a single platform, interface easily with the external input and output and are compatible with established semiconductor nanofabrication processes and integrated-optics material systems. Finally, we outline exciting new applications and research directions.

Principles of SWGs

In the simplest case, an SWG consists of periodically arranged dielectric particles with dimensions much smaller than the wavelength, which form an array of Rayleigh scatterers. For conceptual insight into the optical properties of non-resonant metamaterial structures, a good starting point is the treatment of light propagation through

¹National Research Council Canada, Ottawa, Ontario, Canada. ²Universidad de Málaga, Departamento de Ingeniería de Comunicaciones, ETSI Telecomunicación, Málaga, Spain. ³Bionand Center for Nanomedicine and Biotechnology, Málaga, Spain. ⁴California Institute of Technology, Pasadena, CA, USA. ⁵Duke University, Durham, NC, USA. *e-mail: pavel.cheben@nrc.ca

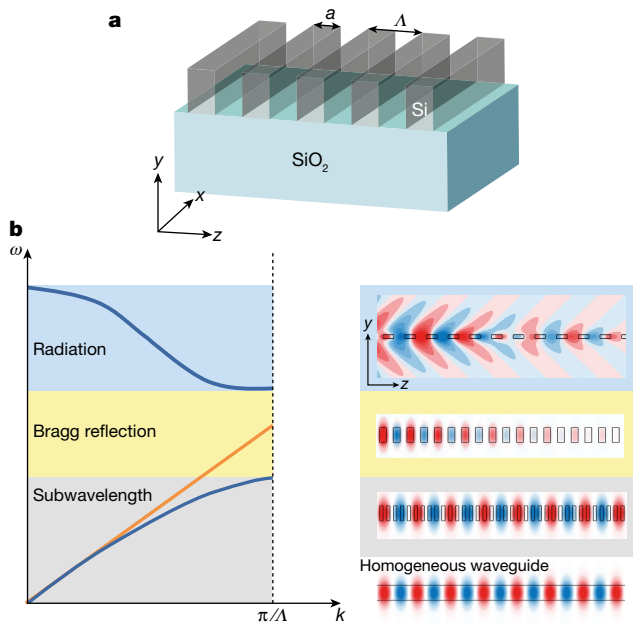


Fig. 1 | Light propagation through a periodic dielectric structure. **a**, Silicon-on-insulator slab waveguide with etched longitudinal or transverse SWG (for light propagation along the z or x axis, respectively). **b**, Schematic dispersion diagram (left) and corresponding electric field profiles (right) of a periodic slab waveguide for the three regimes of subwavelength-guided wave propagation, Bragg reflection and radiation. In the dispersion diagram, the red line is the dispersion of a homogeneous waveguide with an equivalent core refractive index. In the right panel, positive values of the electric field are shown in blue, negative values in red and zero values in white. The black rectangles represent silicon segments.

a finely stratified medium proposed by Rytov¹, where a simple one-dimensional periodic structure consisting of alternating slabs of dielectric materials with refractive indices n_1 and n_2 is considered. It is well known that such a periodic structure can act as a diffraction grating. Rytov found that if the grating period is much smaller than the wavelength of the light, the SWG is optically equivalent to a uniaxial crystal with optic axis perpendicular to the layers. Light incident on the grating can have electric field polarization parallel or perpendicular to the periodic interfaces, and the respective equivalent refractive indices are given by:

$$\begin{aligned} n_{\parallel}^2 &\approx \frac{a}{\Lambda} n_1^2 + \left(1 - \frac{a}{\Lambda}\right) n_2^2 + \mathcal{O}\left(\frac{\Lambda^2}{\lambda^2}\right) \\ n_{\perp}^2 &\approx \frac{a}{\Lambda} n_1^{-2} + \left(1 - \frac{a}{\Lambda}\right) n_2^{-2} + \mathcal{O}\left(\frac{\Lambda^2}{\lambda^2}\right) \end{aligned} \quad (1)$$

Here, a is the width of a slab of material with index n_1 , Λ is the grating period and λ is the free-space wavelength. In the long-wavelength limit, the refractive index approaches a static value with correction terms of the order of Λ^2/λ^2 . This treatment of the grating structure as an equivalent homogeneous material is also referred to as homogenization or effective-medium theory^{2,3,16}. We note that the refractive index of the equivalent homogenous material is polarization-dependent, that is, the material is birefringent.

Over the past decades, fabrication technology has progressed to a point where thin dielectric or metallic films deposited on substrates can be routinely patterned with structures of dimensions that are substantially smaller than the wavelength of the light. As an important example we discuss SWGs etched into silicon-on-insulator wafers for use in integrated photonic circuits; see Fig. 1a. A silicon slab waveguide can be patterned with SWGs of longitudinal, transverse or two-dimensional

periodicity. For periodic longitudinal gratings with periodicity along the axis of propagation, Bragg resonance arises when the period equals the guided half-wavelength, that is, $\Lambda_{\text{Bragg}} = \lambda_{\text{guided}}/2 = \lambda/(2n_{\text{eff}})$, where n_{eff} is the waveguide mode effective index. In general, from photonic crystal theory⁴⁹ it is known that light propagation through a periodic slab waveguide is governed by the dispersion relation shown in Fig. 1b (left). In the diagram, three regimes can be identified: the subwavelength, Bragg and radiation regimes. In the Bragg regime (that is, within the photonic bandgap), no propagating optical mode exists, and a guided wave entering a periodic waveguide in this frequency range decays exponentially within the grating owing to optical reflection. In the radiation regime, the structure acts as a diffraction grating, leading to radiation of the optical power from the waveguide into free space above and below, as seen in Fig. 1b (right). As a consequence of Bloch's theorem, for shorter subwavelength periods, the waveguide (which has discrete translational symmetry) can support localized Floquet–Bloch modes that propagate without loss. The Floquet–Bloch mode is characterized by an electric field that can be expressed along the propagation direction as a plane wave modulated by a periodic amplitude function of the same periodicity as the waveguide. When the grating periodicity is considerably below the wavelength, photonic crystal effects are relatively unimportant. Consistent with effective-medium theory, the structured slab core acts as a homogeneous medium²⁷, which is well approximated as a uniaxial crystal³⁷ with refractive index tensor elements $n_{xx} = n_{yy} = n_{\parallel}$ and $n_{zz} = n_{\perp}$ under the coordinate system defined in Fig. 1a. According to equation (1), by adjusting the filling factor, a/Λ , of the grating, n_{\parallel} and n_{\perp} can be tuned between the refractive indices of the constituent core (Si) and cladding (SiO₂) materials, thereby enabling engineering of the metamaterial refractive index locally on the chip. This is further illustrated in Fig. 1b (left), where the red line shows the dispersion relation of a homogeneous slab waveguide with a core refractive index that results from blending the refractive indices of the constituent materials of the SWG slab waveguide. In the long-wavelength limit (small wavenumber k), the SWG waveguide is optically equivalent to a homogeneous waveguide with an effective core index determined by the filling factor, whereas considerably deviating behaviour is observed for shorter wavelengths approaching the Bragg resonance. Lossless mode propagation is observed not only in the deep-subwavelength regime, but also throughout a transition region of the dispersion diagram towards the photonic bandgap. This is of practical importance because the feature sizes required for an SWG structure in the transition region make it much more amenable to existing fabrication techniques than a deep-subwavelength structure. The ability to control the dispersion and anisotropy of SWG waveguides in the transition region provides a powerful design tool to engineer the wavevectors of the propagating modes (see Box 1). Gratings in the transition region are also used to manipulate free-space beams⁵⁰.

It is important to keep in mind that in k space the transition region of the dispersion diagram is adjacent to the Bragg reflection and radiation regimes, and even small deviations from periodicity that introduce additional spatial frequencies into the subwavelength structure can lead to optical transmission losses by reflection and radiation. Such non-periodicities are introduced through unavoidable fabrication imperfections or by necessity when creating waveguide transitions. For example, great care must be taken in the design of SWG waveguide tapers and transitions to photonic wire waveguides to avoid additional losses that can be incurred by perturbing the periodicity. We expect that limiting radiation losses will become an important practical consideration for photonic components based on transformation optics or inverse design techniques^{33,51,52}, which generally employ non-periodic subwavelength structures.

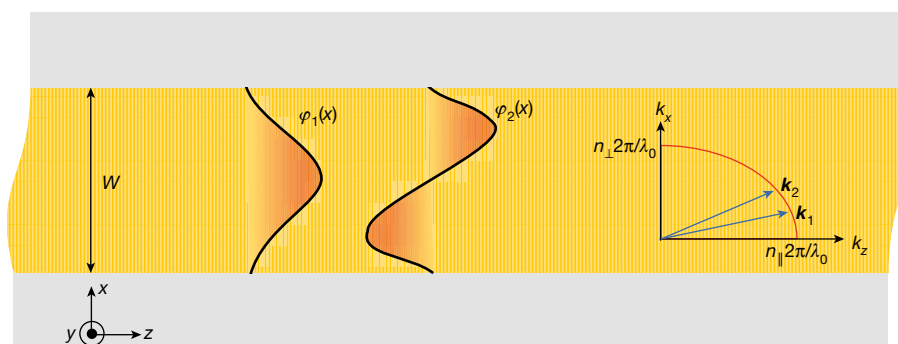
We have described how macroscopic optical material properties, such as birefringence and variable local refractive index profiles, can be artificially generated and engineered by constructing a metamaterial from non-resonant dielectric constituents. In a similar way, creating a metamaterial composed of optically resonant building blocks makes it possible to synthesize artificial bulk materials or surfaces with

BOX I

Waveguiding in an anisotropic material

We consider a multimode waveguide of width W , made of a uniaxial crystal with refractive indices $n_{xx}=n_{\parallel}$ and $n_{zz}=n_{\perp}$ (see figure). The guided modes, φ_m , propagate along the z direction and are polarized in the x direction. For the purpose of illustration, we assume strong guiding, so that the optical modes are confined in the waveguide core with a sinusoidal profile, $\varphi_m(x) \approx \sin(k_{x,m}x)$, where the lateral wavenumber is given approximately by $k_{x,m} \approx m\pi/W$ for the m th guided mode (in the figure, $m \in \{1, 2\}$). The longitudinal component, $k_{z,m}$, of the wavevector yields the mode effective index $n_{\text{eff},m}=k_{z,m}/(2\pi/\lambda)$, which governs phase matching and beating of the waveguide modes and is thus instrumental in the design of integrated devices. From the elliptical dispersion relation of the crystal, $(k_{z,m}/n_{\parallel})^2 + (k_{x,m}/n_{\perp})^2 = (2\pi/\lambda)^2$, and under the paraxial approximation $k_x \ll 2\pi/\lambda$, the mode effective indices are found to be $n_{\text{eff},m} \approx n_{\parallel} - m^2\lambda^2 n_{\parallel}^2 / (8W^2 n_{\perp}^2)$. The filling factor and the period of the grating provide control over n_{\parallel} and n_{\perp} (see ‘Principles of SWGs’) and, consequently, over the effective index and dispersion of the mode.

In devices based on the multimode interference (Talbot self-imaging) effect, the imaging distance is governed by the beat length, L_{π} , of the two lowest-order modes, that is, $L_{\pi} = 2\pi/(k_{z,1} - k_{z,2}) \approx 4W^2 n_{\parallel}^2 / (3\lambda n_{\parallel})$. By engineering the SWG waveguide, the imaging distance can become wavelength-independent, enabling broadband operation³⁷.



interesting and often exotic optical properties. Negative-index materials consisting of arrays of split-ring resonators may be the most prominent example⁶. Although homogenization theories are not strictly applicable to common resonant metamaterial structures owing to the length scales involved, a numerical field-averaging study has shown that an effective-medium picture often provides a useful approximation⁵³. We have encountered a similar situation in practical non-resonant metamaterials: because of fabrication constraints, the metamaterial does not operate in the deep-subwavelength regime, where effective-medium theory is strictly valid, but in the transition region. Unlike the non-resonant subwavelength waveguides used in integrated optics, resonant metamaterials have mostly been implemented in a planar-optics geometry, with light incident on a metasurface from free space. For example, plasmonic nanoantenna arrays on dielectric substrates allow precise control of optical beams⁵⁴. Because metallic materials generally cause appreciable optical losses, alternative lower-loss materials are being explored⁵⁵. There is also surging interest in all-dielectric resonant metasurfaces using Mie resonators as building blocks to achieve effects such as wavefront shaping, optical Huygens surfaces and magnetic mirrors^{14,17,18}. A more detailed discussion of the underlying physical principles of the various resonant metamaterials can be found in a recent review article²². An interesting new concept is the use of these resonant metasurfaces on top of planar waveguides to achieve on-chip optical functions such as mode conversion, polarization rotation and asymmetric transmission⁵⁶, thus opening up the prospect of exploiting the properties of resonant metamaterials in integrated optics.

SWG waveguides and applications

SWG waveguides exploit the ever-improving resolution afforded by complementary metal–oxide–semiconductor (CMOS) lithography techniques, which allow structures with feature sizes below 100 nm to be routinely fabricated in silicon, to locally engineer the material refractive index^{24,28}. The straightforward integration of SWG waveguides with planar silicon-strip waveguides, as illustrated in Fig. 2, has enabled a broad range of integrated optical devices with outstanding performance and growing market relevance. A key factor for the

success of SWG structures is their ease of fabrication alongside standard silicon components, typically using lithography with a single full-etch step. The structural period required for subwavelength operation is $\Lambda < \Lambda_{\text{Bragg}} \approx 300$ nm at telecommunication wavelengths ($\lambda \approx 1.55$ μm). This is well within the range of both electron-beam lithography and wafer-scale deep-ultraviolet lithography, albeit with some limitations in the available filling factors, to comply with the minimum feature sizes of about 50 nm and 100 nm, respectively. For wider (multimode) waveguides with several hundreds of periods, the main fabrication challenge in the short term arises from disorder in the placement of the silicon segments, which changes the translational symmetry of the structure abruptly and must be well below 5 nm to avoid transmission losses⁵⁷. The constraints of minimum feature size and disorder gradually relax for longer wavelengths, making SWG structures particularly promising for the mid-infrared^{14,58,59}.

SWG structures open up unique possibilities of advancing the integration of complex functionalities in silicon chips. A crucial first step in this integration is efficient coupling to optical fibres that link the on-chip device to the exterior system, providing, for instance, medium- and long-haul transmission of information in data- and telecommunication networks. Although the strong light confinement of conventional silicon photonic waveguides allows the realization of compact, tightly integrated photonic circuits, it also hampers direct butt-coupling to optical fibres owing to the large mismatch in mode size, by a factor of roughly 600 for a standard SMF-28 optical fibre. By contrast, mode size can be increased in an SWG waveguide, where light is delocalized from the silicon core as the overall refractive index is reduced (see Fig. 2). Thus, by gradually reducing the filling factor and the width of the SWG waveguide as it approaches the chip edge, the mode size and effective index can be matched to the fibre mode. This yields virtually polarization-independent coupling, which is more difficult to achieve with conventional ‘inverse tapers’⁶⁰. The efficiency exceeds 90% over a bandwidth of more than 100 nm at telecommunication wavelengths for a high-numerical-aperture fibre³⁶.

For coupling to standard fibres, the silicon substrate must be partially removed to avoid leakage of the expanded mode field into the

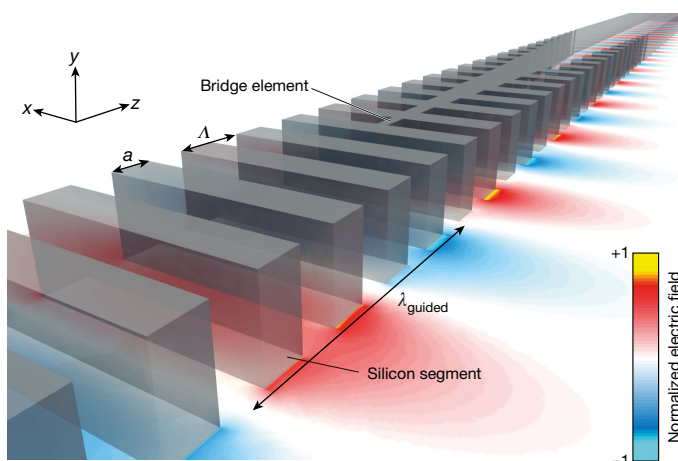


Fig. 2 | Light propagation in a silicon waveguide with an SWG core.

In an SWG waveguide the silicon segments (translucent grey blocks) are spaced with a period, Δ , smaller than the half-wavelength of the guided light wave, $\lambda_{\text{guided}}/2$, so that no diffraction effects arise. Instead, the segmented structure behaves like an anisotropic homogeneous waveguide that blends the refractive indices of the constituent materials, resulting in a reduced mode effective index and an expanded mode size compared to a silicon-strip waveguide. Gradually adding 'bridge' elements in the gaps between the silicon segments provides a nearly lossless transition to the homogeneous silicon waveguide. The colour map shows the normalized electric field of the fundamental horizontally polarized mode.

substrate⁶¹, as demonstrated by IBM researchers⁴³. Such SWG fibre-to-chip couplers can pave the way to efficient, low-cost packaging of silicon photonic chips⁶². An attractive alternative to butt-coupling is offered by surface grating couplers, which operate by diffracting light from the waveguides towards an optical fibre and can thus be placed anywhere on the chip surface. Fully etched grating couplers, apodized with transversal SWGs, have demonstrated peak coupling efficiencies well above 80%⁴⁰, which is considered the threshold for many commercial applications. Without using SWG structures, comparable efficiencies are only achieved with more complex dual-etch-step fabrication processes⁶³. Such processes can also be used in combination with sub-wavelength structures to create perfectly vertical grating couplers⁶⁴ that allow straightforward packaging. Judicious design of the subwavelength structure can even yield polarization-insensitive couplers, illustrated in Fig. 3a, with the additional ability to focus the light in the chip plane⁶⁵. In these grating couplers, the direction of the free-space-diffracted beam is controlled by manipulating its phase profile by introducing local phase changes at the subwavelength scale, as in a metasurface. Therefore, this type of structure can be regarded as a waveguide grating hybridized with an optical metasurface.

A considerable practical constraint of grating couplers is their limited spectral bandwidth of approximately 35 nm (measured at 1 dB, that is, 80% of maximum efficiency) near a wavelength of 1.55 μm , because the momentum-matching requirement in the grating equation imposes a variation of the diffraction angle with wavelength. This variation is proportional to the grating refractive index. By using SWG structures to decrease the index, a 1-dB bandwidth of 90 nm has been demonstrated, albeit at the expense of coupling efficiency⁶⁶. Thus, achieving simultaneous broadband and high-efficiency operation is a challenge. Prism-assisted SWG couplers could potentially provide such a solution³⁸. SWG structures have also been used for coupling light into suspended germanium waveguides at mid-infrared wavelengths⁴⁴, but still with comparatively low efficiencies (around 10%).

Once light is coupled into a nanophotonic waveguide, backscatter arising from the strong interaction of the mode field with the rough sidewalls⁶⁷ can pose a major challenge for reflection-sensitive applications, such as on-chip light sources. The delocalization of the mode in an SWG waveguide can be exploited to diminish this interaction and

reduce backscatter by two orders of magnitude⁶⁸, which may alleviate the need for complex on-chip isolators. Likewise, this reduced interaction with the silicon waveguide core reduces the effective nonlinear coefficient in an SWG waveguide by more than a factor of ten compared to a conventional silicon waveguide, thereby suppressing nonlinear impairments and permitting high-speed data transmission⁶⁹. The same principle enables on-chip time delays of the order of tens of picoseconds by using SWG waveguides of identical length but different group indices, synthesized by changes in the duty cycle⁴¹. It has also been shown that the dispersion profile of such waveguides, with a silicon nitride cladding, can be tailored to obtain both large normal and low anomalous dispersion, which is promising for optical signal processing applications⁷⁰. Furthermore, the periodic nature of the optical field in SWG waveguides (see Fig. 2) creates equally periodic optical forces that can trap nanoparticles both at the sides of the silicon segments and in the gaps between them⁷¹. The working distance for particle trapping is enhanced by the delocalized mode field in SWG waveguides compared to conventional waveguides.

Although SWG structures in the waveguide core produce mode delocalization, the anisotropy of a judiciously designed SWG cladding can effectively enhance modal confinement. Indeed, when a waveguide core made of an isotropic material is embedded in an anisotropic cladding, total internal reflection requires only that the refractive index of the core material be larger than that of the cladding in the direction perpendicular to the propagation. Counter-intuitively, a large refractive index of the cladding in the direction parallel to the propagation will then increase the decay rate of the evanescent field⁷². Such an anisotropic cladding was implemented by subwavelength patterning (parallel to the direction of propagation) of the waveguide material around the silicon core to demonstrate reduced crosstalk between densely packed waveguides⁷³. SWG claddings patterned perpendicular to the direction of propagation are advantageously used for waveguides operating in the mid-infrared, where the silicon dioxide layer that optically insulates the waveguide core from the silicon substrate becomes lossy. The gaps in the SWG cladding allow the removal of the lossy oxide layer using hydrofluoric acid, resulting in suspended waveguides that are laterally supported by the SWG segments⁵⁹. Using this approach, silicon waveguides with losses less than 1 dB cm^{-1} at $\lambda = 3.8 \mu\text{m}$ and 3 dB cm^{-1} at $\lambda = 7.7 \mu\text{m}$, as well as slotted waveguides with losses of 8 dB cm^{-1} at $\lambda = 2.3 \mu\text{m}$, have been fabricated^{58,59,74}.

On-chip devices and systems

Devices for on-chip beam splitting, polarization management and spectral filtering are essential building blocks for integrated optical systems, and SWG structures are facilitating key advances in all three areas. Directional couplers are widely used to implement integrated beam splitters. However, their operation principle, which is based on the interference of a pair of supermodes in two parallel waveguides, offers a limited operational bandwidth (about 25 nm at telecommunication wavelengths). Superimposing an SWG structure on a conventional directional coupler provides control over the dispersion of these supermodes and enables operation over a bandwidth of around 100 nm^{75,76}. Even broader bandwidths, in excess of 500 nm, can be obtained through the Talbot (self-imaging) effect in multimode SWG waveguides (see Fig. 3b), resulting in a threefold enhancement of the bandwidth compared to conventional devices³⁷. This is achieved by taking advantage of the SWG anisotropy to attain a wavelength-independent imaging distance, as outlined in Box 1. Extending this device to four inputs and four outputs, while maintaining excess losses and imbalance below 1 dB, would yield a telecommunication quadrature hybrid with a bandwidth of several hundreds of nanometres. When fabricated with wafer-scale lithography, such a device would enable the production of optical coherent receiver systems covering several optical communication bands at once.

By building on the concept of topology optimization^{77,78}, extremely compact beam splitters, with a footprint smaller than $3 \mu\text{m} \times 3 \mu\text{m}$, can be achieved using intricate subwavelength structures obtained by

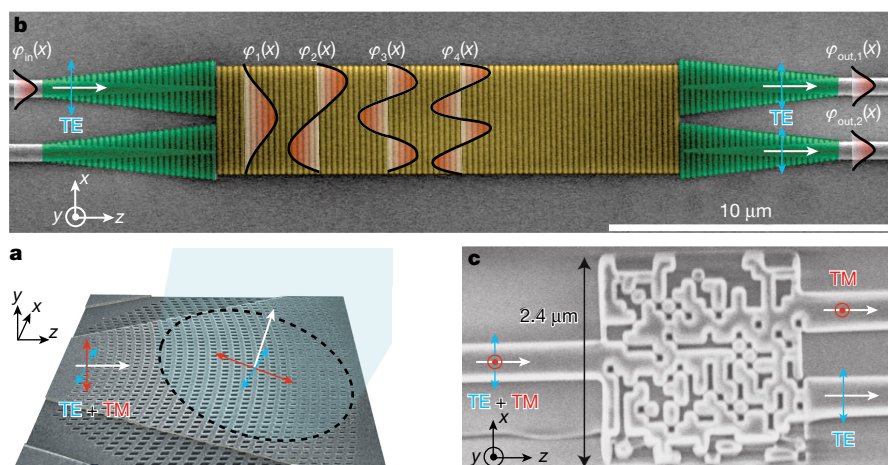


Fig. 3 | Subwavelength engineered waveguide devices for fibre-to-chip coupling, beam splitting and polarization splitting. **a**, A focusing, polarization-independent fibre-to-chip grating coupler. Light is coupled from an optical fibre (shown in blue) into the chip (x - z plane) through a diffraction grating along the z direction on the chip surface. The grating is curved to provide focusing of the light beam in the chip plane. The SWG (oriented along the x direction) provides control over the amplitude and phase of the diffracted field, thereby enabling operation at both polarizations, along the x (transverse electric, TE) and y (transverse magnetic, TM) directions. Figure adapted from ref. ⁶⁵, Optical Society of America. **b**, A broadband on-chip beam splitter based on the multimode interference (Talbot) effect. The input mode, $\varphi_{in}(x)$, travels in a silicon-

wire waveguide and is gradually transformed to a wider SWG waveguide mode (green area). At the abrupt transition to the multimode SWG waveguide (yellow area), several higher-order modes are excited and interfere as they propagate, forming images of the expanded input mode. Coupling these images to the output modes ideally yields $\varphi_{out,1} = \varphi_{in}/\sqrt{2}$ and $\varphi_{out,2} = i\varphi_{in}/\sqrt{2}$. By exploiting the anisotropy of the multimode SWG waveguide, the imaging distance can be made almost wavelength-independent, thereby achieving broadband operation (see Box 1). Polarization is transverse electric, that is, in the plane of the chip, along the x direction. **c**, An ultra-compact polarization beam splitter based on a numerically optimized nanopattern of subwavelength 'pixels' that create a metamaterial. Figure adapted from ref. ⁵², Springer Nature Ltd.

numerical minimization techniques, albeit with a more limited bandwidth of about 60 nm⁷⁹. Similar numerical approaches have been used to design ultra-compact devices for on-chip polarization management. One example is a polarization splitter, shown in Fig. 3c, with a footprint of only $2.4\mu\text{m} \times 2.4\mu\text{m}$ and an extinction ratio of 10 dB over a 30 nm bandwidth⁵². This performance is still limited compared to that of polarization splitters based on bent directional couplers, which offer extinction ratios in excess of 25 dB over a comparable bandwidth but are also about six times longer⁸⁰. Polarization rotation with an extinction ratio of 10 dB, insertion losses of 2 dB and a very competitive 140 nm bandwidth has been recently reported in a 4- μm -long device designed using genetic algorithms⁸¹. A single device that functions as a polarization splitter with a polarization rotator at one of its outputs has been realized using phase matching between the vertically polarized mode of a silicon wire waveguide and the horizontally polarized mode of an SWG waveguide^{82,83}. The device achieves a remarkable tolerance to fabrication deviation of up to ± 40 nm, whereas many conventional devices tolerate only errors of the order of ± 10 nm. Thus, compact, practical SWG-based polarization splitters and rotators with extinction ratios above 20 dB and sub-decibel losses with bandwidths over 100 nm seem within reach in the near future^{42,84}.

For applications in on-chip spectral filtering, Bragg gratings based on the same principle of successive constructively interfering reflections as their fibre-optic counterparts⁸⁵ are commonly used. However, in fully etched nanophotonic waveguides, it is challenging to achieve the low reflection coefficients and long grating lengths required for filtering bandwidths below a few nanometres. This limitation can be overcome by using a waveguide with two corrugations interleaved at the subwavelength scale⁸⁶, which yields a bandwidth of around 1 nm with a resonance depth of 40 dB. Such small bandwidths could previously be achieved in silicon waveguides only in dual-etch-depth designs⁸⁷. Hybrid SWG-Bragg spectral filters with even smaller bandwidths of about 100 pm have recently been proposed⁸⁸. Other structures of interest are contra-directional couplers, which are based on phase-matching modes propagating in two parallel waveguides in different directions via a grating. These couplers offer a wide free spectral range for add-drop wavelength multiplexing but suffer from undesired codirectional

coupling. Using an SWG waveguide in one of the coupler arms promotes contra-directional coupling while producing a strong phase mismatch that efficiently suppresses the codirectional coupling⁸⁹.

System-level integration of SWG structures, while still at an early stage, is already showing outstanding results. Compact Fourier-transform interferometers that synthesize optical path differences using SWG waveguides have been shown to achieve spectral resolution of 50 pm at near-infrared wavelengths⁹⁰. Grating couplers based on two-layer nanostructures and with 92% efficiency have been fabricated using a standard CMOS process⁹¹, paving the way for system-level integration of electronics and photonic nanostructures⁴⁸.

An outstanding challenge in integrated photonics is achieving dynamic control of the coupling between guided waves and free-space propagating beams. Encouraging results have been reported on waveguide phased arrays⁹², including the first demonstration of coherent solid-state light detection and ranging (LIDAR) using optical phased arrays in a silicon photonics platform⁹³. Recent advances in the surging field of optical metasurfaces^{16,21,22,54} have also opened prospects for bridging this gap. While the SWG structures that we have discussed typically control the behaviour of light during propagation in the two-dimensional chip plane, the third spatial dimension can be accessed by integrating a metasurface directly on a planar waveguide circuit. This can enable dynamic control of free-space beams emitted off-chip for agile interfacing of integrated optical devices with the external environment. Tuning of the overall metasurface response can be achieved using many different physical mechanisms²². Although a planar waveguide circuit with an integrated dynamic metasurface has not yet been demonstrated, several promising candidates have been reported. Independent electrical modulation of both amplitude and phase has been demonstrated, enabling electrical switching of diffracted beams at high frequencies (more than 10 MHz)⁹⁴. In this structure, tunability arises from field-effect modulation of the complex refractive index of the conducting oxide layers incorporated into metasurface antenna elements. Applying an electrical bias between metal and indium tin oxide (ITO) changes the sign of the real part of the dielectric permittivity of ITO. When the relative dielectric permittivity, ϵ_r , of ITO is in the epsilon-near-zero region ($-1 < \epsilon_r < 1$), a large

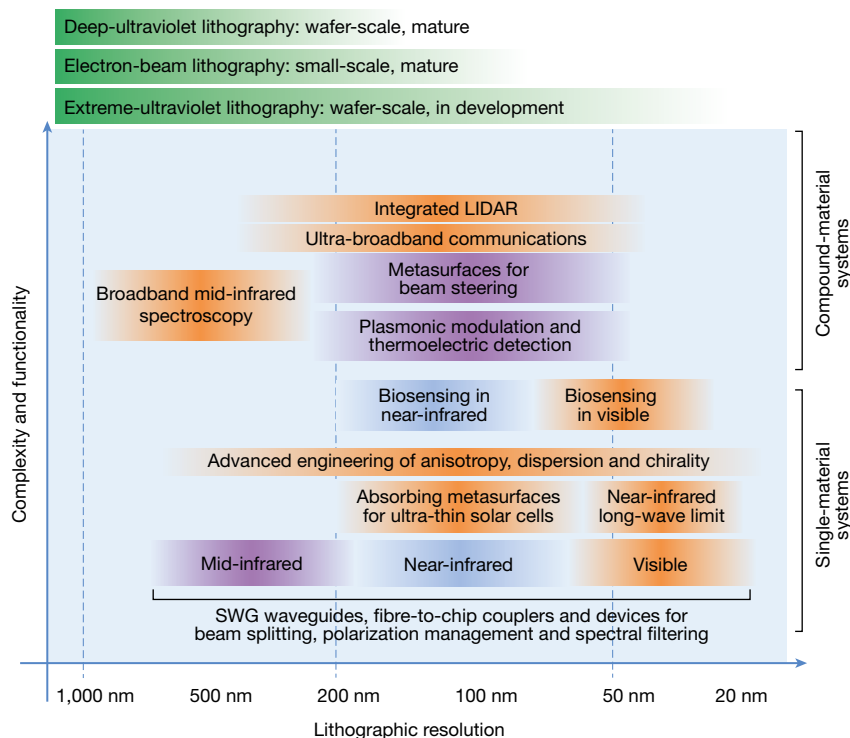


Fig. 4 | A roadmap for integrated SWG metamaterial devices and systems. Blue-shaded boxes indicate devices and systems that have already been demonstrated, whereas orange-shaded boxes refer to expected future implementations. Purple boxes show systems for which substantial progress has been made but no waveguide-integrated validation is currently available (beam steering with metasurfaces, plasmonic modulation and thermoelectric detection). In the mid-infrared, some of the functionalities shown (waveguides, fibre-to-chip couplers, beam

splitting) have been implemented, but others have not (polarization management, spectral filtering). The range of feature sizes that can be synthesized with different lithography techniques is indicated by the green bars. We differentiate between systems that can be implemented using a single material (typically silicon) and systems that require additional materials for light generation, detection or active tuning. See 'Conclusions and outlook' for a complete description.

electric-field enhancement occurs in the accumulation layer for near-infrared wavelengths, providing an efficient way to electrically modulate the optical phase and amplitude, with high modulation speed and low power consumption. Active metasurfaces have also been explored for electrostatic control of the scattered field phase in the mid-infrared. With active control of phase, one can engineer arbitrary phase fronts in both space and time, enabling dynamically reconfigurable metasurface devices. Electrostatic-phase control in the mid-infrared using graphene- and ITO-integrated resonant structures has demonstrated tunabilities of 55° at $7.7\ \mu\text{m}$ (graphene)⁹⁵ and 180° at $5.95\ \mu\text{m}$ (ITO)⁹⁶. Recently, a widely tunable phase modulation in excess of 230° was demonstrated using an electrostatically gate-tunable graphene-gold metasurface⁹⁷ at $8.5\ \mu\text{m}$.

Nanomechanical devices actuated by thermal, electrostatic, magnetic and optical effects can also impact future integrated photonic technologies. Several proof-of-principle demonstrations of nonlinear, switching, electro-optical and magneto-optical functionalities in nanomechanical devices have shown growing potential for practical device integration. This emerging field has recently been reviewed elsewhere²⁰.

Conclusions and outlook

SWG-integrated structures enable the development of a rapidly growing range of high-performance devices at near-infrared telecommunication wavelengths^{36,37,40,81,83,86}. The incorporation of these all-dielectric components into more complex, planar waveguide architectures and CMOS processes is expected to continue^{68,91,98}, whereas immersion lithography techniques⁹⁹ will further facilitate their mass fabrication and commercial exploitation. Some of the SWG structures shown in Fig. 4 have already been successfully brought into the burgeoning field of integrated mid-infrared photonics, including low-loss SWG-engineered waveguides⁵⁸ and grating couplers⁴⁴, whereas

others are expected to follow in the near future. In addition, improved lithographic resolution achieved by extreme-ultraviolet techniques¹⁰⁰ will facilitate the use of SWG structures at visible wavelengths and will open up the long-wave limits in the near- and mid-infrared. The flexibility and dispersion-less nature of SWG structures in this regime makes them ideal for the implementation of transformation optics¹⁰. Furthermore, superior lithographic resolutions would enable the development of SWG-enhanced biosensors in the visible wavelength range, where some of the most sensitive devices reported until now operate¹⁰¹. The anisotropic^{37,72} and dispersive⁷⁰ properties of subwavelength nanostructures, which have barely been explored, offer further research routes in all wavelength ranges. In combination with compound material systems that enable bandgap-free photodecay¹⁰² and photodetector integration¹⁰³, low-loss SWG waveguides⁵⁸ and devices could provide on-chip spectroscopy systems⁹⁰ in the mid-infrared fingerprint region, with applications in environmental monitoring and security. Integrated coherent receivers for ultrabroad optical communications are also becoming feasible, with broadband fibre-to-chip couplers already available³⁶, and broadband polarization management^{81–83} and optical quadrature hybrids³⁷ within reach. Future on-chip integration of agile metasurfaces reconfigurable at high speeds⁹⁴ is envisioned to allow the development of integrated coherent phased arrays at visible and infrared frequencies⁹³, enabling functions such as electronic beam steering and focusing, which have previously been available only in microwave RADAR systems.

Received: 18 October 2017; Accepted: 13 June 2018;

Published online 29 August 2018.

1. Rytov, S. M. Electromagnetic properties of a finely stratified medium. *Sov. Phys. JETP* **2**, 466–475 (1956).
2. Mait, J. N. & Prather, D. W. (eds) *Selected Papers on Subwavelength Diffractive Optics* (SPIE Optical Engineering Press, Bellingham, 2001).

3. Lalanne, P., Astilean, S., Chavel, P., Cambriil, E. & Launois, H. Design and fabrication of blazed binary diffractive elements with sampling periods smaller than the structural cutoff. *J. Opt. Soc. Am. A* **16**, 1143–1156 (1999).
4. Bloembergen, N. & Sievers, A. J. Nonlinear optical properties of periodic laminar structures. *Appl. Phys. Lett.* **17**, 483–486 (1970).
5. van der Ziel, J. P. Phase-matched harmonic generation in a laminar structure with wave propagation in the plane of the layers. *Appl. Phys. Lett.* **26**, 60–61 (1975).
6. Smith, D. R., Padilla, W. J., Vier, D. C., Nemat-Nasser, S. C. & Schultz, S. Composite medium with simultaneously negative permeability and permittivity. *Phys. Rev. Lett.* **84**, 4184–4187 (2000).
7. Pendry, J. B. Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
8. Walser, R. M. Electromagnetic metamaterials. *Proc. SPIE* **4467**, <https://doi.org/10.1117/12.432921> (2001).
9. Shelby, R. A., Smith, D. R. & Schultz, S. Experimental verification of a negative index of refraction. *Science* **292**, 77–79 (2001).
10. Pendry, J. B. Controlling electromagnetic fields. *Science* **312**, 1780–1782 (2006).
11. Fedotov, V. A., Schwanecke, A. S., Zheludev, N. I., Khardikov, V. V. & Prosvirnin, S. L. Asymmetric transmission of light and enantiomerically sensitive plasmon resonance in planar chiral nanostructures. *Nano Lett.* **7**, 1996–1999 (2007).
12. Yen, T. J. et al. Terahertz magnetic response from artificial materials. *Science* **303**, 1494–1496 (2004).
13. Urbas, A. M. et al. Roadmap on optical metamaterials. *J. Opt.* **18**, 093005 (2016).
14. Kuznetsov, A. I., Miroshnichenko, A. E., Brongersma, M. L., Kivshar, Y. S. & Luk'yanchuk, B. Optically resonant dielectric nanostructures. *Science* **354**, aag2472 (2016).
15. Zhu, A. Y., Kuznetsov, A. I., Luk'yanchuk, B., Engheta, N. & Genevet, P. Traditional and emerging materials for optical metasurfaces. *Nanophotonics* **6**, 452–471 (2017).
16. Lalanne, P. & Chavel, P. Metalenses at visible wavelengths: past, present, perspectives. *Laser Photonics Rev.* **11**, 1600295 (2017).
17. Staude, I. & Schilling, J. Metamaterial-inspired silicon nanophotonics. *Nat. Photon.* **11**, 274–284 (2017).
18. Jahani, S. & Jacob, S. All-dielectric metamaterials. *Nat. Nanotechnol.* **11**, 23–36 (2016).
19. Zheludev, N. I. Obtaining optical properties on demand. *Science* **348**, 973–974 (2015).
20. Zheludev, N. I. & Plum, E. Reconfigurable nanomechanical photonic metamaterials. *Nat. Nanotechnol.* **11**, 16–22 (2016).
21. Genevet, P. & Capasso, F. Holographic optical metasurfaces: a review of current progress. *Rep. Prog. Phys.* **78**, 024401 (2015).
22. Chen, H.-T., Taylor, A. J. & Yu, N. A review of metasurfaces: physics and applications. *Rep. Prog. Phys.* **79**, 076401 (2016).
23. Vivien, L. & Pavesi, L. (eds) *Handbook of Silicon Photonics* (CRC Press, Boca Raton, 2013).
24. Cheben, P., Xu, D.-X., Janz, S. & Densmore, A. Subwavelength waveguide grating for mode conversion and light coupling in integrated optics. *Opt. Express* **14**, 4695–4702 (2006).
- This paper proposed SWG metamaterial structures for silicon-strip waveguides.**
25. Schmid, J. H. et al. Gradient-index antireflective subwavelength structures for planar waveguide facets. *Opt. Lett.* **32**, 1794–1796 (2007).
- This study demonstrated SWG structures in a silicon-on-insulator rib waveguide.**
26. Bock, P. J. et al. Subwavelength grating periodic structures in silicon-on-insulator: a new type of microphotonic waveguide. *Opt. Express* **18**, 20251 (2010).
27. Halir, R. et al. Waveguide sub-wavelength structures: a review of principles and applications. *Laser Photonics Rev.* **9**, 25–49 (2015).
28. Cheben, P. et al. Refractive index engineering with subwavelength gratings for efficient microphotonic couplers and planar waveguide multiplexers. *Opt. Lett.* **35**, 2526–2528 (2010).
- This study demonstrated refractive-index-engineered SWG silicon waveguide devices.**
29. Bock, P. J. et al. Subwavelength grating crossings for silicon wire waveguides. *Opt. Express* **18**, 16146–16155 (2010).
30. Levy, U. et al. Inhomogeneous dielectric metamaterials with space-variant polarizability. *Phys. Rev. Lett.* **98**, 243901 (2007).
- This paper reported on refractive-index engineering with SWG structures in slab waveguides.**
31. Valentine, J., Li, J., Zentgraf, T., Bartal, G. & Zhang, X. An optical cloak made of dielectrics. *Nat. Mater.* **8**, 568–571 (2009).
32. Hunt, J. et al. Planar, flattened Luneburg lens at infrared wavelengths. *Opt. Express* **20**, 1706 (2012).
33. Gabrielli, L. H. & Lipson, M. Transformation optics on a silicon platform. *J. Opt.* **13**, 024010 (2011).
34. Zentgraf, T., Valentine, J., Tapia, N., Li, J. & Zhang, X. An optical 'Janus' device for integrated photonics. *Adv. Mater.* **22**, 2561–2564 (2010).
35. Glesk, I. et al. All-optical switching using nonlinear subwavelength Mach-Zehnder on silicon. *Opt. Express* **19**, 14031 (2011).
36. Cheben, P. et al. Broadband polarization independent nanophotonic coupler for silicon waveguides with ultra-high efficiency. *Opt. Express* **23**, 22553–22563 (2015).
37. Halir, R. et al. Ultra-broadband nanophotonic beamsplitter using an anisotropic sub-wavelength metamaterial. *Laser Photonics Rev.* **10**, 1039–1046 (2016).
- This study exploited the anisotropy of SWG structures to achieve broadband operation.**
38. Sánchez-Postigo, A. et al. Broadband fiber-chip zero-order surface grating coupler with 0.4 dB efficiency. *Opt. Lett.* **41**, 3013–3016 (2016).
39. Wangüemert-Pérez, J. G. et al. Evanescent field waveguide sensing with subwavelength grating structures in silicon-on-insulator. *Opt. Lett.* **39**, 4442–4445 (2014).
- This paper proposed the use of SWG for enhanced waveguide sensing.**
40. Benedikovic, D. et al. Subwavelength index engineered surface grating coupler with sub-decibel efficiency for 220-nm silicon-on-insulator waveguides. *Opt. Express* **23**, 22628–22635 (2015).
41. Wang, J. et al. Subwavelength grating enabled on-chip ultra-compact optical true time delay line. *Sci. Rep.* **6**, 30235 (2016).
42. Xu, Y. & Xiao, J. Ultracompact and high efficient silicon-based polarization splitter-rotator using a partially-etched subwavelength grating coupler. *Sci. Rep.* **6**, 27949 (2016).
43. Barwicz, T. et al. A metamaterial converter centered at 1490nm for interfacing standard fibers to nanophotonic waveguides. In *Proc. Optical Fiber Communication Conference M21.3* (Optical Society of America, 2016).
44. Kang, J. et al. Focusing subwavelength grating coupler for mid-infrared suspended membrane germanium waveguides. *Opt. Lett.* **42**, 2094–2097 (2017).
45. Benedikovic, D. et al. L-shaped fiber-chip grating couplers with high directionality and low reflectivity fabricated with deep-UV lithography. *Opt. Lett.* **42**, 3439–3442 (2017).
46. Flueckiger, J. et al. Sub-wavelength grating for enhanced ring resonator biosensor. *Opt. Express* **24**, 15672–15686 (2016).
47. Yan, H. et al. Unique surface sensing property and enhanced sensitivity in microring resonator biosensors based on subwavelength grating waveguides. *Opt. Express* **24**, 29724–29733 (2016).
- This work demonstrated enhanced surface sensitivity for SWG waveguides.**
48. Sun, C. et al. Single-chip microprocessor that communicates directly using light. *Nature* **528**, 534–538 (2015).
49. Joannopoulos, J. D., Johnson, S. G., Winn, J. N. & Meade, R. D. *Photonic Crystals: Molding the Flow of Light* 2nd edn (Princeton University Press, Princeton, 2008).
50. Chang-Hasnain, C. & Yang, W. High-contrast gratings for integrated optoelectronics. *Adv. Opt. Photonics* **4**, 379–440 (2012).
51. Piggott, A. Y. et al. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nat. Photon.* **9**, 374–377 (2015).
52. Shen, B., Wang, P., Polson, R. & Menon, R. An integrated-nanophotonics polarization beamsplitter with $2.4 \times 2.4 \mu\text{m}^2$ footprint. *Nat. Photon.* **9**, 378–382 (2015).
53. Smith, D. R. & Pendry, J. B. Homogenization of metamaterials by field averaging (invited paper). *J. Opt. Soc. Am. B* **23**, 391 (2006).
54. Kildishev, A. V., Boltasseva, A. & Shalae, V. M. Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013).
55. Boltasseva, A. & Atwater, H. A. Low-loss plasmonic metamaterials. *Science* **331**, 290–291 (2011).
56. Li, Z. et al. Controlling propagation and coupling of waveguide modes using phase-gradient metasurfaces. *Nat. Nanotechnol.* **12**, 675–683 (2017).
57. Ortega-Moñux, A. et al. Disorder effects in subwavelength grating metamaterial waveguides. *Opt. Express* **25**, 12222–12236 (2017).
58. Penadés, J. S. et al. Suspended silicon waveguides for long-wave infrared wavelengths. *Opt. Lett.* **43**, 795–798 (2018).
59. Penadés, J. S. et al. Suspended silicon mid-infrared waveguide devices with subwavelength grating metamaterial cladding. *Opt. Express* **24**, 22908–22916 (2016).
60. Almeida, V. R., Panepucci, R. R. & Lipson, M. Nanotaper for compact mode conversion. *Opt. Lett.* **28**, 1302 (2003).
61. Sarmiento-Merenguel, J. D. et al. Controlling leakage losses in subwavelength grating silicon metamaterial waveguides. *Opt. Lett.* **41**, 3443–3446 (2016).
62. Barwicz, T. et al. A novel approach to photonic packaging leveraging existing high-throughput microelectronic facilities. *IEEE J. Sel. Top. Quant.* **22**, 455–466 (2016).
- This study used a metamaterial fibre-chip coupler for IBM's advanced photonic packaging.**
63. Mekis, A. et al. A grating-coupler-enabled CMOS photonics platform. *IEEE J. Sel. Top. Quant.* **17**, 597–608 (2011).
64. Watanabe, T., Ayata, M., Koch, U., Fedoryshyn, Y. & Leuthold, J. Perpendicular grating coupler based on a blazed anti-back-reflection structure. *J. Light Technol.* **35**, 4663–4669 (2017).
65. Cheng, Z. & Tsang, H. K. Experimental demonstration of polarization-insensitive air-cladding grating couplers for silicon-on-insulator waveguides. *Opt. Lett.* **39**, 2206–2209 (2014).
66. Wang, Y. et al. Design of broadband subwavelength grating couplers with low back reflection. *Opt. Lett.* **40**, 4647–4650 (2015).
67. Melati, D., Melloni, A. & Morichetti, F. Real photonic waveguides: guiding light through imperfections. *Adv. Opt. Photonics* **6**, 156 (2014).
68. Peng, B. et al. Metamaterial waveguides with low distributed backscattering in production O-band Si photonics. In *Proc. Optical Fiber Communication Conference Tu3K.3* (Optical Society of America, 2017).
- This work showed the low-backscatter advantage of SWG waveguides.**

69. Gao, G. et al. Transmission of 2.86 Tb/s data stream in silicon subwavelength grating waveguides. *Opt. Express* **25**, 2918 (2017).
 70. Benedikovic, D. et al. Dispersion control of silicon nanophotonic waveguides using sub-wavelength grating metamaterials in near- and mid-IR wavelengths. *Opt. Express* **25**, 19468–19478 (2017).
 71. Ma, K., Han, S., Zhang, L., Shi, Y. & Dai, D. Optical forces in silicon subwavelength-grating waveguides. *Opt. Express* **25**, 30876–30884 (2017).
 72. Jahani, S. & Jacob, Z. Transparent sub-diffraction optics: nanoscale light confinement without metal. *Optica* **1**, 96–100 (2014).
 73. Jahani, S. et al. Controlling evanescent waves on-chip using all-dielectric metamaterials for dense photonic integration. *Nat. Commun.* **9**, 1893 (2018).
 74. Zhou, W. et al. Fully suspended slot waveguides for high refractive index sensitivity. *Opt. Lett.* **42**, 1245–1248 (2017).
 75. Halir, R., Cheben, P., Xu, D., Schmid, J. H. & Janz, S. Colorless directional coupler with dispersion engineered sub-wavelength structure. *Opt. Express* **20**, 13470–13477 (2012).
 76. Wang, Y. et al. Compact broadband directional couplers using subwavelength gratings. *IEEE Photonics J.* **8**, 1–8 (2016).
 77. Jensen, J. S. & Sigmund, O. Topology optimization for nano-photonics. *Laser Photonics Rev.* **5**, 308–321 (2011).
 78. Frandsen, L. H. et al. Topology optimized mode conversion in a photonic crystal waveguide fabricated in silicon-on-insulator material. *Opt. Express* **22**, 8525 (2014).
 79. Lu, L. et al. Inverse-designed single-step-etched colorless 3 dB couplers based on RIE-lag-insensitive PhC-like subwavelength structures. *Opt. Lett.* **41**, 5051–5054 (2016).
 80. Wu, H. & Dai, D. High-performance polarizing beam splitters based on cascaded bent directional couplers. *IEEE Photonics Technol. Lett.* **29**, 474–477 (2017).
 81. Yu, Z., Cui, H. & Sun, X. Genetic-algorithm-optimized wideband on-chip polarization rotator with an ultrasmall footprint. *Opt. Lett.* **42**, 3093–3096 (2017).
 82. Xiong, Y. et al. Polarization splitter and rotator with subwavelength grating for enhanced fabrication tolerance. *Opt. Lett.* **39**, 6931–6934 (2014).
 83. He, Y. et al. Silicon polarization splitter and rotator using a subwavelength grating based directional coupler. In *Proc. Optical Fiber Communication Conference Th1G.6* (Optical Society of America, 2017).
 84. Xu, L. et al. Polarization beam splitter based on MMI coupler with swg birefringence engineering on soi. *IEEE Photonics Technol. Lett.* **30**, 403–406 (2018).
 85. Kashyap, R. *Fiber Bragg Gratings* 2nd edn (Academic Press, Burlington, 2009).
 86. Pérez-Galacho, D. et al. Optical pump-rejection filter based on silicon sub-wavelength engineered photonic structures. *Opt. Lett.* **42**, 1468–1471 (2017).
 87. Wang, X. et al. Narrow-band waveguide Bragg gratings on SOI wafers with CMOS-compatible fabrication process. *Opt. Express* **20**, 15547 (2012).
 88. Čtyroký, J. et al. Design of narrowband Bragg spectral filters in subwavelength grating metamaterial waveguides. *Opt. Express* **26**, 179–194 (2018).
 89. Naghdi, B. & Chen, L. R. Silicon photonic contradirectional couplers using subwavelength grating waveguides. *Opt. Express* **24**, 23429–23438 (2016).
 90. Podmore, H. et al. Demonstration of a compressive-sensing Fourier-transform on-chip spectrometer. *Opt. Lett.* **42**, 1440–1443 (2017).
 91. Notaros, J. et al. Ultra-efficient CMOS fiber-to-chip grating couplers. In *Proc. OFC 2016* 1–3 (IEEE, 2016).
 92. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).
 93. Poulton, C. V. et al. Coherent solid-state LIDAR with silicon photonic optical phased arrays. *Opt. Lett.* **42**, 4091 (2017).
 94. Huang, Y. W. et al. Gate-tunable conducting oxide metasurfaces. *Nano Lett.* **16**, 5319–5325 (2016).
- This work demonstrates a gate-tunable metasurface that allows dynamic electrical control of light phase and amplitude, with a modulation frequency greater than 10 MHz.**
95. Dabidian, N. et al. Experimental demonstration of phase modulation and motion sensing using graphene-integrated metasurfaces. *Nano Lett.* **16**, 3607–3615 (2016).
 96. Park, J., Kang, J. H., Kim, S. J., Liu, X. & Brongersma, M. L. Dynamic reflection phase and polarization control in metasurfaces. *Nano Lett.* **17**, 407–413 (2017).
 97. Sherrott, M. C. et al. Experimental demonstration of >230° phase modulation in gate-tunable graphene-gold reconfigurable mid-infrared metasurfaces. *Nano Lett.* **17**, 3027–3034 (2017).
 98. Barwicz, T., Kamlapurkar, S., Martin, Y., Bruce, R. L. & Engelmann, S. A silicon metamaterial chip-to-chip coupler for photonic flip-chip applications. In *Proc. Optical Fiber Communication Conference Th2A.39* (Optical Society of America, 2017).
 99. Jeong, S. et al. Low-loss, flat-topped and spectrally uniform silicon-nanowire-based 5th-order CROW fabricated by ArF-immersion lithography process on a 300-mm SOI wafer. *Opt. Express* **21**, 30163–30174 (2013).
 100. McGrath, D. ASML claims major EUV milestone. *EETimes* http://www.eetimes.com/document.asp?doc_id=1332012 (2017).
 101. Gavela, A. F., García, D. G., Ramirez, J. C. & Lechuga, L. M. Last advances in silicon-based optical biosensors. *Sensors* **16**, 285 (2016).
 102. Mauser, K. W. et al. Resonant thermoelectric nanophotonics. *Nat. Nanotechnol.* **12**, 770–775 (2017).
 103. Berini, P. Surface plasmon photodetectors and their applications. *Laser Photonics Rev.* **8**, 197–220 (2014).
- Acknowledgements** We are grateful to S. Janz, D.-X. Xu, A. Ortega-Moñux, Í. Molina-Fernández, J. G. Wangüemert-Pérez, J. Lapointe, J. Čtyroký, C. Alonso-Ramos, D. Benedikovic, G. Mashanovich, A. V. Velasco, W. Ye, M. L. Calvo, L. Vivien, Y. Grinberg, D. Melati and M. Dado for discussions. R.H. acknowledges financial support from Ministerio de Economía y Competitividad, Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad (cofinanciado FEDER) Proyecto TEC2016-80718-R. H.A.A. acknowledges financial support from the Air Force Office of Scientific Research under grant number FA9550-16-1-0019.
- Author contributions** J.H.S., R.H., P.C. and H.A.A. wrote the manuscript. P.C. and D.R.S. contributed to its preparation.
- Competing interests** The authors declare no competing interests.
- Additional information**
- Reprints and permissions information** is available at <http://www.nature.com/reprints>.
- Correspondence and requests for materials** should be addressed to P.C.
- Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The emergence of Zika virus and its new clinical syndromes

Theodore C. Pierson^{1*} & Michael S. Diamond^{2,3,4,5*}

Zika virus (ZIKV) is a mosquito-transmitted flavivirus that has emerged as a global health threat because of its potential to generate explosive epidemics and ability to cause congenital disease in the context of infection during pregnancy. Whereas much is known about the biology of related flaviviruses, the unique features of ZIKV pathogenesis, including infection of the fetus, persistence in immune-privileged sites and sexual transmission, have presented new challenges. The rapid development of cell culture and animal models has facilitated a new appreciation of ZIKV biology. This knowledge has created opportunities for the development of countermeasures, including multiple ZIKV vaccine candidates, which are advancing through clinical trials. Here we describe the recent advances that have led to a new understanding of the causes and consequences of the ZIKV epidemic.

ZIKV was first isolated from a febrile sentinel monkey in Uganda in 1947. Serological data suggest that ZIKV was distributed widely throughout Africa and subsequently in Asia despite the absence of described morbidity¹. The first ZIKV outbreak to garner international attention occurred on Yap Island in the Western Pacific Ocean in 2007. Forty-nine confirmed human cases were reported². More than half of the inhabitants of Yap were believed to have been infected, with many experiencing rash, fever and arthralgia. ZIKV activity was next detected in the islands of French Polynesia in 2013, with a larger number of infections. Some of the unique clinical features of ZIKV (for example, Guillain-Barré syndrome, congenital malformations and the presence of the virus in semen) were identified during this outbreak or later in retrospective studies³. ZIKV was introduced in Brazil in late 2013 or early 2014, spread rapidly within the northeast part of the country, and was repeatedly introduced into regions of the Americas⁴. The large number of infections and links to congenital neurodevelopmental defects identified this epidemic as an international public health emergency. ZIKV activity in the Americas peaked in the early spring of 2016, followed by a marked decrease in reported cases in 2017, which is probably attributable to the effect of herd immunity. Seroprevalence studies suggest that 63% of the inhabitants of Salvador, Brazil were infected during this outbreak⁵. By 2017, more than 220,000 confirmed and 580,000 suspected cases were reported in 52 countries or territories in the Americas (PAHO Zika Cumulative Cases; 4 January 2018).

Viral structure

Flaviviruses encapsulate a positive-stranded RNA genome, which encodes a single open reading frame flanked by two structured untranslated regions (UTRs) (Fig. 1). The single viral polypeptide is processed by host and viral proteases into three structural proteins (capsid (C), pre-membrane (prM) and envelope (E)) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5), the latter of which mediate genome replication, viral polypeptide processing and modulation of the host response. ZIKV virions are composed of the three structural proteins, a lipid envelope and the viral RNA genome. The E protein is an elongated protein, which consists of three ectodomains connected by short flexible loops and is anchored to the

viral membrane by a helical stem and two antiparallel transmembrane domains (Fig. 1). In most ZIKV strains, the E protein is modified by a single N-linked glycan at position E154 located in domain I (E-DI); some pre-epidemic ZIKV strains from Africa lack this N-linked glycan and are less neuroinvasive⁶. Although ZIKV is most closely related to another African flavivirus called Spondweni virus (approximately 68% E protein amino acid identity), it shares sequence similarity with other flaviviruses. Approximately 50% of the E protein is conserved among ZIKV and dengue (DENV) virus strains. Although conserved regions could be targeted by broadly reactive protective antibodies^{7,8}, this feature complicates the development of virus-specific diagnostics, and raises the prospect of adverse immune reactions in individuals exposed sequentially to ZIKV and DENV⁹. Two lineages of ZIKV (African and Asian) differ from each other by approximately 10% at the nucleotide level¹. ZIKV strains in the Americas descend from the Asian lineage¹⁰.

The prM protein is a small glycoprotein that is connected to the viral membrane by antiparallel transmembrane helices. Immature ZIKV virions assemble on endoplasmic reticulum-derived membranes as particles on which trimers of prM-E heterodimers form spiked projections arranged with icosahedral symmetry¹¹. These non-infectious virions transit through the secretory pathway and undergo a conformational change upon exposure to the mildly acidic environment in the trans-Golgi network that enables cleavage of prM by a host furin-like protease. Mature virions are characterized by a relatively smooth structure created by 90 E protein dimers orientated parallel to the plane of the viral membrane^{12,13} (Fig. 1). However, prM cleavage can be inefficient, resulting in partially mature infectious virions that retain uncleaved prM. Although the efficiency of prM cleavage on ZIKV relative to other flaviviruses is not yet known, partially mature ZIKV virions may be infectious. Although there has been rapid progress into the structural biology of ZIKV virions, these studies capture only a single state of mature and immature virions. Insights into the structural ensembles that are adopted by the virus throughout its replication cycle await further study¹⁴. Additional comparative analyses of prM-E proteins and virion structures of pre- and post-epidemic ZIKV strains could provide further insights into how ZIKV evolved to become more pathogenic and cause distinct clinical syndromes.

¹Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ²Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. ³Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA. ⁴Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, USA. ⁵Andrew M. and Jane M. Bursky Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, St. Louis, MO, USA. *e-mail: piersontc@mail.nih.gov; diamond@wustl.wustl.edu

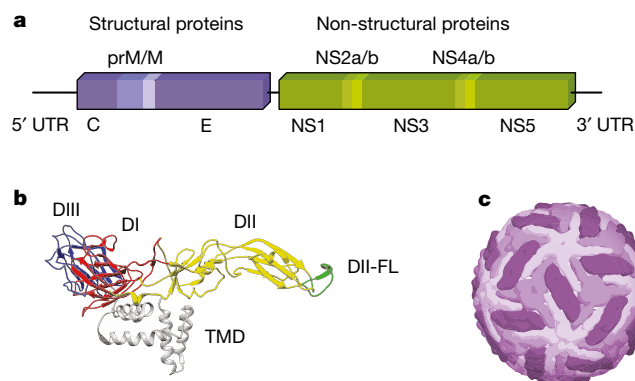


Fig. 1 | Genomic organization and structure of ZIKV. **a**, ZIKV encapsulates a positive-stranded genomic RNA sequence that encodes a single polyprotein, which is cleaved into three structural and seven non-structural proteins. **b**, E consists of three ectodomains (E-DI, E-DII and E-DIII, which are shown in red, yellow and blue, respectively) and is anchored into the viral membrane by two anti-parallel transmembrane domains (TMD; grey). A highly conserved fusion loop (DII-FL) is located at the distal end of DII (green). **c**, Mature virions incorporate 180 copies of the E protein arranged with icosahedral symmetry. E proteins are in three environments defined by their proximity to the two-, three- and fivefold symmetry axes (shades of purple)^{12,13}.

ZIKV infection in different clinical contexts

The clinical syndrome caused by ZIKV in humans was historically reported as a mild influenza-like illness that resolved within days and occurred in approximately 20% of infected individuals² (Fig. 2). However, in French Polynesia, the rate of symptomatic infections was higher (approximately 50%)¹⁵. The most common signs and symptoms of ZIKV infection in the French Polynesian and American outbreaks occurred within 3–7 days of being bitten by a mosquito and included fever (72%), arthralgia and myalgia (65%), conjunctivitis (63%), headache (46%), fatigue and/or rash³. During the recent epidemics, ZIKV infection also has been associated, albeit infrequently, with severe disease in adults, including multi-organ failure¹⁶, meningitis and encephalitis¹⁷, and thrombocytopenia¹⁸. Although ZIKV generally does not cause fatal disease in adults, mortality has been described in children with sickle cell disease, adults with cancer¹⁶ and those cases who develop Guillain–Barré syndrome¹⁹, a progressive polyneuropathy linked to ZIKV infection, which occurred in 1/6,500 to 1/17,000 individuals in endemic regions^{3,20} (Fig. 2).

A distinguishing feature of ZIKV infection during the recent epidemic is an apparent broadening of cellular tropism and persistence in multiple organs; this has resulted in ostensibly new clinical manifestations. It remains unclear whether this reflects a fundamental change in ZIKV virulence or whether this is now appreciated owing to the greater number of diagnosed infections. ZIKV persists in whole-blood and immune-sanctuary sites. In multiple case reports, whole blood from non-pregnant adults remained positive for ZIKV RNA for 60–100 days, long after serum and other body fluids became negative^{21,22}. In a pediatric study that evaluated the acute phase of infection, ZIKV principally infected CD14⁺CD16⁺ monocytes in the blood²³. ZIKV can replicate persistently within cells of the anterior and posterior chambers of the eye²⁴, which causes conjunctivitis, maculopathy and uveitis, the latter of which can result in blindness^{25,26}. ZIKV persistence in the eye has been detected in mice and humans for up to 30 days^{24,27}, and is speculated to be a means of direct transmission^{16,24}.

Another site of ZIKV persistence is the male reproductive tract. Persistent ZIKV RNA in sperm and semen has been reported in humans for months²⁸, although infectious virus was limited to the first few weeks after disease onset²⁹. Experiments in monkeys also show persistence of ZIKV in male reproductive tract tissues³⁰. Studies in mice have demonstrated that ZIKV can replicate in cells of the testis, including spermatogonia, Sertoli cells and Leydig cells, which results in destruction of testicular architecture, reduction in sperm counts, lower

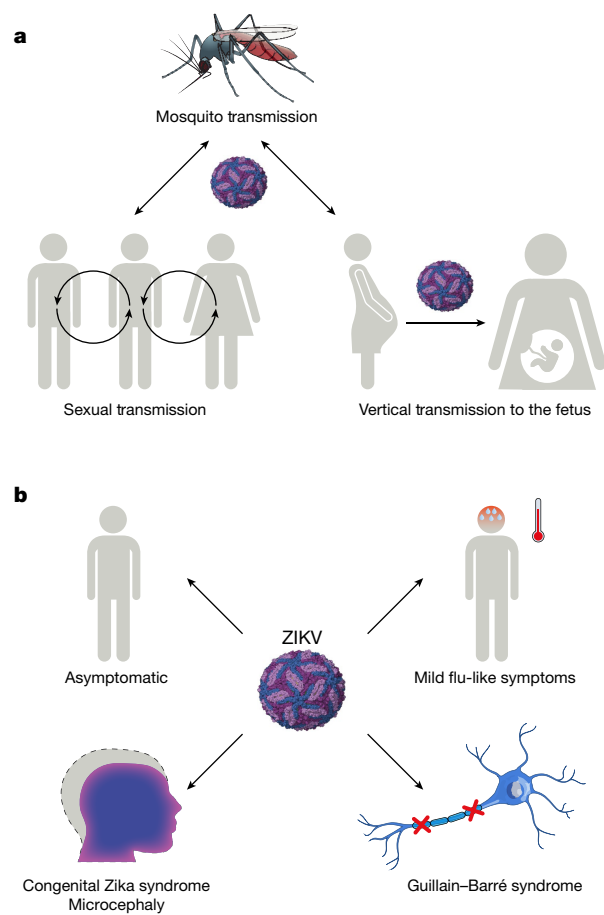


Fig. 2 | Transmission and clinical manifestations of ZIKV.

a, Transmission. ZIKV is transmitted in an epidemic cycle between *Aedes* mosquitoes and humans. ZIKV can also be transmitted through sexual contact or vertical transmission from an infected pregnant mother to her fetus. **b**, Clinical syndromes. Most ZIKV infections are asymptomatic. Among symptomatic cases, most patients develop an illness characterized by fever, rash, conjunctivitis, headaches and muscle and/or joint pain. During pregnancy, ZIKV infection can result in microcephaly, congenital ZIKV syndrome (CZS) and fetal demise. In a subset of adults, infection is linked to Guillain–Barré syndrome, which can result in muscle weakness and paralysis.

levels of sex hormones and reduced fertility^{31,32}. Oligospermia and haematospermia have been observed in humans after ZIKV infection and are speculated to affect fertility³³. The high viral load in seminal fluid can lead to sexual transmission from men-to-women³⁴ and men-to-men³⁵ (Fig. 2).

ZIKV is linked to the development of Guillain–Barré syndrome in a small percentage of adults^{36,37}, although causality has not been proven. Guillain–Barré syndrome is an acute inflammatory immune-mediated polyneuropathy that typically presents with paresthesia, weakness and pain, but can progress to paralysis and even death. Case reports have been published of ZIKV-associated Guillain–Barré syndrome in French Polynesia in 2013 and in the Americas^{38,39}. Although more investigations into its cause are needed, leading hypotheses include B or T cell-mediated immunopathology due to viral antigen mimicry or direct viral infection and injury of cells of the peripheral nervous system.

Considerable effort has been made to define why ZIKV has teratogenic abilities. Initial studies focused on the placenta, because it acts as a structural and immunologic barrier between the maternal uterine-derived decidua and the developing fetus during pregnancy⁴⁰ (Fig. 3). The maternal–fetal interface is characterized by an apposition of maternal decidua with fetally derived proliferative cytotrophoblasts and terminally differentiated syncytiotrophoblasts, the latter of which create a multi-nucleated cell barrier. In the first trimester, the

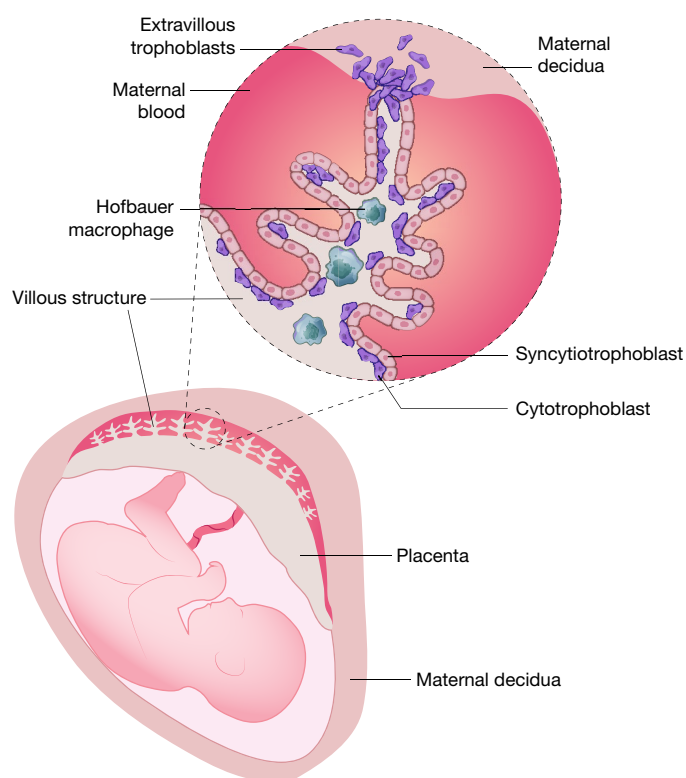


Fig. 3 | An ineffective placental barrier to congenital ZIKV infections. Vertical transmission requires ZIKV to spread to the immunologically privileged fetus. Maternal–fetal interactions occur at placental villous structures that are in contact with maternal blood and the adjacent decidua. Placental villi are lined by syncytiotrophoblasts that form a physical barrier against infection and have active roles in antiviral immunity through production of IFN λ and microRNAs^{43,147}. ZIKV targets several placental cells including trophoblasts, Hofbauer macrophages and endothelial cells.

human placenta becomes haemochorial, directly contacting maternal blood, which allows for the exchange of gases, nutrients and wastes. Syncytiotrophoblasts also function as a physical and immunological barrier between the maternal and fetal circulation to prevent spread of microbial agents. The maternal–fetal interface is defined by the differentiation of floating and anchoring villous structures. The core of the villous structure beneath the syncytiotrophoblast layer consists of fetal Hofbauer macrophages, fibroblasts and endothelial cells.

The importance of trophoblasts to fetal ZIKV infection was first identified in cell culture studies. Human trophoblast cells isolated early during pregnancy propagated ZIKV to high levels^{41,42}. However, human fetal-derived chorionic villi or trophoblast cells obtained later during gestation supported less ZIKV infection, which probably reflects pregnancy stage-dependent effects of trophoblast differentiation and immunological maturation⁴³. The reduced susceptibility to ZIKV infection at later gestational stages results in part from distinct innate immune profiles, including production or response to type-III interferon- λ (IFN λ)^{43,44} or differential expression of putative entry receptors. ZIKV can also replicate in placental cytotrophoblasts or primitive trophoblasts^{41,42}, fetal Hofbauer macrophages⁴⁵ and fetal endothelial cells⁴⁶. Infection of fetal Hofbauer macrophages may be particularly important in humans, as ZIKV RNA and antigen have been localized to these cells in the placentas of women who had pregnancy losses during the first or second trimester⁴⁷. These data suggest that ZIKV may be transmitted to fetuses by a transplacental route. This tropism of ZIKV may not be unique among flaviviruses, as inoculation of human placental explants or pregnant mice with the related West Nile and Powassan viruses also resulted in infection of trophoblasts and injury to the placenta⁴⁸.

Thousands of infants have been born with ZIKV-induced neurologic sequelae (CZS) that will impair their future neurodevelopmental

function⁴⁹. Insight into the molecular basis of CZS has been informed by experiments with cultured neuroprogenitor stem cells, cortical organoids, brain slices and mouse brain inoculation studies⁵⁰. In these models, ZIKV preferentially targets progenitor cells of the cerebral cortex and results in reduced cell proliferation and differentiation, and increased inflammation and increased cell death⁵¹. ZIKV may also infect microglia in the brain, which can produce inflammatory cytokines (for example, TNF, IL-6 and IL-1 β) that inhibit proliferation and differentiation of neuronal precursor cells⁵². Studies have also suggested that ZIKV can infect and modulate immune responses of astrocytes, in a manner that depends on expression of the cell surface protein AXL^{53,54}.

There is a range of penetrance of CZS in different geographical regions for reasons that remain unclear. In one case series of 182 symptomatic, ZIKV-infected pregnant women in Brazil, a remarkable 42% of fetuses had abnormal clinical or brain imaging findings, and adverse outcomes were noted regardless of trimester of infection⁵⁵. Retrospective analysis in French Polynesia also found an increased risk of microcephaly associated with ZIKV infection⁵⁶. In a study of ZIKV infection during pregnancy in the territories of the United States, 5% of fetuses or infants had birth defects, with deleterious outcomes occurring in all trimesters⁵⁷. Neuroimaging of the brains of congenitally infected neonates have reported hypoplasia of the cerebellum and brainstem, ventriculomegaly, myelination defects, calcifications and cortical malformations. The clinical presentation of CZS is variable and includes fetal demise and microcephaly⁵⁵, as well as sensorineural hearing loss, ocular abnormalities and arthrogryposis (joint contractures). Severe microcephaly is associated with mental retardation, learning disabilities, behavioural abnormalities, muscle weakness and altered muscle tone. The devastating effects of CZS are reflected by data from one longitudinal cohort of infants that was followed for eight months after birth: 85% had irritability, 56% had altered muscle tone and movement, 50% had epileptic seizures, 15% had dysphagia, and all of these infants had abnormal brain imaging studies⁵⁸. In a second study of CZS-affected children from Brazil who were followed for 19–24 months, most had severe motor impairment, seizure disorders, hearing and vision abnormalities and sleep difficulties; this resulted in these children falling far behind in achieving developmental milestones, indicating the need for long-term support⁵⁹.

Key questions related to CZS remain: (1) What is the risk of vertical transmission and disease in mothers who contract an asymptomatic ZIKV infection? (2) For neonates of mothers infected with ZIKV during the different stages of pregnancy who appear normal at birth, what is the risk of neurodevelopmental disorders? (3) How does pre-existing maternal flavivirus immunity impact pregnancy-associated disease?

Pathogenesis models in animals

There has been remarkable progress establishing mouse models of ZIKV infection and disease. Peripheral challenge of adult wild-type mice with ZIKV isolates results in little virus replication and no disease⁶⁰. In comparison, infection of neonatal mice often causes non-fatal yet severe neurological disease characterized by tremors, ataxia, seizures and microcephaly. Because of the failure of immunocompetent adult mice to sustain a ZIKV infection, several groups evaluated the capacity of mice with IFN signalling deficiencies to support ZIKV replication, as this strategy had been used for other flaviviruses (for example, DENV and yellow fever virus). Mice that lacked the type-I *Ifnar1* gene, or both *Ifnar1* and the type-II *Ifngr* receptors, or that received a neutralizing anti-*Ifnar1* antibody were susceptible to infection by most strains of ZIKV, and this resulted in central nervous system disease and death following inoculation through several different routes.

Mouse models of ZIKV infection during pregnancy have been developed to study transmission, teratogenicity and vaccine protection. Subcutaneous inoculation of pregnant *Ifnar1*^{−/−} or wild-type dams treated with anti-*Ifnar1* antibodies resulted in ZIKV infection of trophoblasts in the placenta, which enabled transplacental transmission, spread to the fetal brain and fetal demise⁴¹. The gestational

stage of the dam affects clinical outcome: ZIKV infection during early pregnancy (embryonic day (E)6) resulted in placental insufficiency and fetal demise, infections at mid-pregnancy (E9) resulted in reduced cranial dimensions consistent with microcephaly, and infection during late pregnancy (E12) caused no apparent fetal disease⁴⁴. These results correlate with human studies, which show ZIKV-associated microcephaly occurs more commonly when infections happen during the first and early second trimesters⁶¹.

Pregnant immunocompetent mice do not transmit ZIKV to the placenta or fetus when the virus is inoculated subcutaneously, presumably because of the failure to evade type-I IFN immunity⁴¹. However, intravenous inoculation of pregnant mice with high doses of ZIKV caused cortical brain malformations and fetal ocular abnormalities^{62,63}. As an alternative model, direct ZIKV inoculation into the uterine wall of pregnant mice resulted in placental infection and inflammation, reduced neonatal brain cortical thickness and reduced fetal viability⁶⁴. One study showed that direct viral infection of the fetus is not essential for demise, as placental pathology may be a stronger contributor to adverse pregnancy outcomes⁶⁵. Other groups have injected ZIKV directly into the developing fetus in the cerebroventricular space; this resulted in decreased brain size, thinning of cortical layers, reduced numbers of cortical neural progenitors and neuronal cell death⁶⁶. Intravaginal transmission of ZIKV during pregnancy has also been modelled in *Ifnar1*^{-/-} mice during the progesterone-high, diestrous phase of the estrous cycle⁶⁷. Finally, a recent study generated an immunocompetent mouse model of ZIKV infection and placental transmission by introducing the human *STAT2* gene into the mouse *Stat2* locus⁶⁸.

In Africa, ZIKV undergoes a sylvatic phase with infection of non-human primates (NHPs) occurring in an endemic cycle. NHP models have advantages over rodent models, as ZIKV naturally targets primates and, thus, is more likely to overcome species-specific immune barriers to infection. Moreover, NHPs are evolutionary closer to humans than rodents, and the findings on pathogenesis and host restriction therefore are probably more relevant. Consequently, experimental ZIKV infection in rhesus, cynomolgus and pigtail macaques as well as marmosets has been used to evaluate ZIKV biology^{69–72}. Because there is concern that ZIKV could establish a sylvatic cycle in the Americas, more recent studies have evaluated ZIKV infection in New World marmosets⁷³.

Experimental inoculation of rhesus macaques with ZIKV resulted in weight loss, elevated body temperature, rash and mild hepatitis. These animals developed viraemia that peaks within the first week and then becomes undetectable by day 10, and viral RNA was detected in urine, saliva, lacrimal fluid, cerebrospinal fluid, seminal fluid and vaginal secretions^{69,70}. ZIKV infected multiple tissues of rhesus and cynomolgus macaques including lymphoid organs, the male reproductive tract, the intestines, and the brain and spinal cord^{70,72,74}. Because infected rhesus macaques also developed ZIKV-specific humoral and cell-mediated immune responses^{69–71,74}, this model has been used to study sequential ZIKV and DENV infections and vaccine efficacy^{75,76}.

Although challenging, ZIKV infection of pregnant NHPs is important because the haemochorial placenta and gestational development are more similar to humans than mice. Pregnant rhesus macaques infected with ZIKV developed viraemia that lasted for 30–55 days⁶⁹, which is similar to viraemia observed in pregnant women⁷⁷. The first NHP model of in utero transmission was established after inoculation of a pigtail macaque with an Asian ZIKV strain⁷⁸. Infection resulted in reduced growth of the fetal brain, white matter gliosis and axonal damage, and ZIKV RNA was detected in the placenta, fetal brain and liver. In other studies, pregnant rhesus macaques infected with an Asian or American ZIKV strain also resulted in prolonged maternal viraemia^{79,80}. Fetal head growth in the last month of gestation was decreased, and ZIKV RNA was detected in fetal tissues at birth. Pathological analysis showed neutrophil infiltration at the maternal–fetal interface and brain lesions in fetuses, including microcalcifications, haemorrhage, vasculitis and apoptosis of neuroprogenitor cells^{79,80}. Because 26% of NHPs infected with ZIKV during early gestation experienced fetal demise despite showing few clinical signs, pregnancy loss due to

asymptomatic infection may be underrecognized⁸¹. Vertical transmission in NHP models may provide a platform for testing vaccines and antibody-based therapeutics^{82,83}. In other studies, postnatal ZIKV infection was associated with abnormalities in brain structure, function and behaviour in infant macaques⁸⁴.

Innate immune responses to ZIKV infection

Although the initial innate immune events following ZIKV infection are beginning to be characterized, paradigms for recognition and control by the cytoplasmic (RIG-I-like receptors) and endosomal (Toll-like receptors) viral RNA sensors and signalling through downstream adaptor molecules and transcription factors have been extrapolated largely from studies with other flaviviruses.

Type-I and type-III IFNs induce antiviral states through induction of IFN-stimulated genes (ISG) that control viral replication. Type-I IFNs (for example, IFN α and IFN β) bind to their heterodimeric receptor (IFNAR1/IFNAR2) and promote phosphorylation of JAK1 and TYK2. This activates STAT1 and STAT2 to bind IRF9 and form the IFN-stimulated gene factor 3 complex (ISGF3), which transcriptionally activates hundreds of ISGs. Type-III IFN λ binds to a selectively expressed, heterodimeric receptor (IFNLR1/IL10R3), and analogously promotes ISGF3 nuclear translocation and ISG induction. In addition, IFN λ has antiviral functions against ZIKV in the maternal decidua and placenta during pregnancy^{43,44}. Constitutive secretion of IFN λ by syncytiotrophoblasts correlated with their ability to restrict ZIKV infection⁴³. The importance of IFN signalling in mediating host restriction of ZIKV was shown by the pathogenicity of ZIKV in *Ifnar1*^{-/-} and *Stat2*^{-/-} but not immunocompetent mice^{60,85,86}. Several antiviral effector genes induced by type-I and type-III IFNs reportedly have antiviral effects against ZIKV. Members of the IFITM family and their interacting proteins inhibit ZIKV infection at an entry step in the viral life cycle^{87,88}. Expression of viperin also inhibited ZIKV replication in cells⁸⁹.

ZIKV evades IFN responses by impairing induction and signalling pathways at multiple steps. Human dendritic cells can be infected productively by ZIKV⁹⁰, but do not secrete pro-inflammatory cytokines or type-I IFN, probably because antiviral pathogen-recognition receptors and their downstream signalling pathways are downregulated or evaded⁹¹. ZIKV targets the IFN signalling pathways by inhibiting JAK1 and STAT activity. ZIKV NS5 targets human STAT2, but not mouse Stat2 for proteasomal degradation^{92,93}. In addition, ZIKV infection prevents STAT1 phosphorylation⁹⁰. ZIKV NS1 and NS4B also appear to inhibit IFN β induction at the level of TBK1 activation, and the ZIKV NS2B–NS3 protease impairs IFNAR induction and signalling pathways by targeting human STING but not mouse Sting for cleavage⁹⁴ and by degrading JAK1⁹⁵. Finally, ZIKV NS4B induces elongation of mitochondria, which physically contact the membranes associated with the endoplasmic reticulum that are sites of replication. This restructuring attenuates RIG-I-dependent activation of IFN responses. Beyond viral protein-mediated evasion mechanisms, ZIKV also generates a subgenomic viral RNA that antagonizes RIG-I-induced type-I IFN responses⁹⁶.

Apart from active innate immune evasion mechanisms, ZIKV targets some cells that are inherently deficient in innate immune responses. Primary neural progenitor cells have a delayed innate response to ZIKV infection⁸⁹ and glioblastoma cancer stem cells, which are highly permissive to ZIKV infection, show an absence of IFN signatures⁹⁷. Analogously, in the vagina, ZIKV replication induces a weak antiviral IFN response⁹⁸.

Adaptive immune responses to ZIKV infection

During primary infection, anti-ZIKV IgM becomes detectable as early as three days after onset of illness with most individuals developing responses by day 8. This early antibody response originates from extra-follicular ZIKV-specific plasmablasts, which comprise a large fraction of the circulating B cells^{99,100}. This plasmablast response, however, is transient and lasts only a few weeks¹⁰⁰, with germinal centre-derived

plasma cells starting to produce antibody at this time. Neutralizing antibodies develop within the first week of illness, and as the IgG response matures, inhibitory antibodies in sera accumulate and neutralize virus strains from both Asian and African lineages¹⁰¹.

The functional quality and antigenic targets of ZIKV-induced B cell responses have been evaluated¹⁰². Prior flavivirus immunity is associated with serological cross-reactivity after ZIKV infection^{99,103,104}. In humans with prior DENV immunity, a substantial proportion of anti-ZIKV antibodies generated during acute infection targets the highly conserved fusion loop in E-DII. Plasmablasts from acutely ZIKV-infected, DENV-immune individuals exhibited high levels of somatic hypermutation, with many derived from common memory B cell clones⁹⁹. By contrast, plasmablasts from ZIKV-infected, flavivirus-naïve individuals exhibited less somatic hypermutation or clonal expansion⁹⁹, and antibody responses were more ZIKV-specific¹⁰⁵. In general, cross-reactive antibodies had poorer neutralizing capacity *in vitro* and limited protective activity *in vivo* against ZIKV^{106,107}. Prior flavivirus immunity triggers cross-reactive responses because the memory B cells formed during the first flavivirus infection encounter conserved epitopes present on ZIKV antigens¹⁰². The magnitude and durability of the cross-reactive response may depend on the duration separating the two flavivirus infections and the number of prior exposures^{108,109}.

Functional and structural studies have revealed epitopes on all domains of the ZIKV E protein for engagement by highly neutralizing monoclonal antibodies¹⁰². One class, which consists of antibodies that are cross-reactive with DENV and recognize the quaternary envelope dimer epitope¹¹⁰, neutralized ZIKV infection with high potency⁷ and protected mice and NHPs from ZIKV infection^{8,83} or transplacental transmission⁸. A second class of highly neutralizing and protective anti-ZIKV monoclonal antibodies binds to residues within the lateral ridge epitope of E-DIII and blocks infection at a post-attachment step¹¹¹. E-DIII antibodies appear important for controlling ZIKV, as their depletion from human serum resulted in reduced neutralizing activity against ZIKV¹¹². A third class of ZIKV monoclonal antibodies also protects against vertical transmission of ZIKV in mice¹⁰⁴. The only described monoclonal antibody of this class, ZIKV-117, recognizes an epitope across neighbouring E protein dimers and probably prevents the conformational changes required for pH-dependent fusion in the endosome. Finally, neutralizing monoclonal antibodies binding to additional sites within E-DI and E-DII have also been reported¹¹³.

We are beginning to understand T cell responses against ZIKV. In mice, polyfunctional, cytotoxic CD8⁺ T cells become activated¹¹⁴ and can reduce ZIKV burden, whereas their depletion or genetic absence resulted in greater ZIKV infection and mortality¹¹⁵. Consistent with this observation, adoptive transfer of ZIKV-immune CD8⁺ T cells can protect against ZIKV infection¹¹⁶. Another study identified human-relevant ZIKV CD8⁺ T cell epitopes in naïve and DENV-experienced HLA-transgenic mice and demonstrated that both ZIKV-specific and ZIKV–DENV cross-reactive CD8⁺ T cells can protect against ZIKV infection¹¹⁷. Thus, CD8⁺ T cells probably have a protective activity against ZIKV. Nevertheless, CD8⁺ T cells could have pathological consequences in the brain¹¹⁸ and result in ZIKV-associated paralysis¹¹⁹.

Less is known about human T cell responses to ZIKV. In one study, ZIKV-infected patients developed polyfunctional CD4⁺ T cell responses that produced antiviral cytokines¹⁰⁵. In another study, ZIKV infection induced CD4⁺ T effector cells with a low frequency of IFN- γ production¹²⁰. In comparison, whereas CD8⁺ T cells were highly activated during the viraemic phase (15% to >25% of blood CD8⁺ T cells), low levels of antigen-specific CD8⁺ T cells were identified¹⁰⁵, although others have reported tetramer-positive ZIKV-specific CD8⁺ T cells in blood¹⁰⁰. Another study found that memory T cell responses elicited by prior infection or vaccination with DENV were restimulated with ZIKV-derived peptides¹²¹; the consequence of this expanded cross-reactive response (beneficial or detrimental) remains to be determined. Almost 60% of the ZIKV-specific CD8⁺ T cell response was directed

against the structural proteins¹²², which could be beneficial for vaccines that exclusively target the structural proteins (see Figs. 1, 4).

Possible explanations for the emergence of ZIKV

How ZIKV has changed to cause massive epidemics, congenital defects, infection in immune sanctuary sites, sexual transmission and Guillain–Barré syndrome remains an area of intensive study. Multiple factors may be responsible for the changing epidemiology and disease pathogenesis.

The presence of NS1 in human blood facilitates flavivirus acquisition by mosquito vectors because this protein suppresses the immune functions of the mosquito midgut. A single alanine to valine (A188V) substitution in NS1 of the epidemic ZIKV strains facilitated greater infectivity in *Aedes aegypti* mosquitoes and enzootic transmission¹²³. Clinical isolates from the Americas with a valine at position 188 had higher NS1 antigenemia in mice and were more infectious in mosquitoes than pre-epidemic strains. This same mutation promotes the binding of NS1 to TBK1, resulting in reduced levels of TBK1 phosphorylation and IFN- β expression in human cells⁹⁵. Thus, sequence changes in NS1 during the pre-epidemic to epidemic transition appear to have facilitated immune evasion and enhanced ZIKV transmissibility from hosts to vectors, which creates conditions for epidemic transmission.

Sequence changes have also been hypothesized to affect ZIKV pathogenicity and explain its tropism for fetal neuroprogenitor cells¹⁰. A single serine to asparagine substitution (S139N) in the viral polyprotein (residue 17 of prM) increased ZIKV infectivity in neural progenitor cells that resulted in more severe microcephaly and higher mortality rates in neonatal mice¹²⁴. Evolutionary analysis indicated that the S139N substitution arose just before the 2013 outbreak in French Polynesia and has been maintained during the American epidemic. The mechanistic basis for how the S139N in prM mediates this effect remains uncertain.

Noncoding sequence adaptations that affect the RNA structure may contribute to neuropathogenicity of epidemic strains. One group identified a Musashi protein binding element in the SL2 stem loop of the 3' UTR, with sequence changes between ancestral and contemporary strains immediately upstream of this site¹²⁵. Because Musashi proteins regulate mRNA translation and can modulate progenitor cell growth and differentiation, this group postulated that nucleotide polymorphisms in the 3' UTR might be linked to alterations in Musashi protein binding activity and neurovirulence. A second group showed that Musashi-1 interacts directly with ZIKV genomic RNA and facilitates viral replication¹²⁶. ZIKV infection disrupted the binding of Musashi-1 to its endogenous targets, which deregulated expression of factors that have been implicated in neural stem cell function.

A feature of DENV pathogenesis is that antibodies to one serotype can exacerbate infection with a second serotype via antibody-dependent enhancement (ADE)⁹. ADE occurs when cross-reactive, non-neutralizing quantities of antibodies bind to a heterologous DENV serotype and facilitate infection of myeloid cells that express Fc- γ receptors. Because of the structural similarity between ZIKV and DENV, antibodies produced against these flaviviruses can cross-react^{103,127}. ADE between DENV and ZIKV has been proposed to contribute to the severity of ZIKV disease in the Americas, since the epidemics occurred in regions in which most people are DENV-immune. Indeed, cross-reactive anti-DENV antibodies can enhance ZIKV infection in cell culture^{106,127}. Notwithstanding this observation, ADE in cell culture has been demonstrated for many viruses without evidence of worsened disease in humans. One recent study showed that passive transfer of immune plasma against DENV or West Nile virus can enhance ZIKV infection and pathogenesis in *Stat2*^{-/-} mice¹²⁸, which suggest that pre-existing anti-flavivirus immunity can promote ZIKV pathogenesis; a caveat to this model is that these mice lacked immune T cells, which can limit the effects of ADE in the context of DENV infection. However, no differences in ZIKV infection were observed after inoculation of naïve and flavivirus-immune rhesus macaques⁷⁵. By contrast, prior exposure to ZIKV enhanced DENV

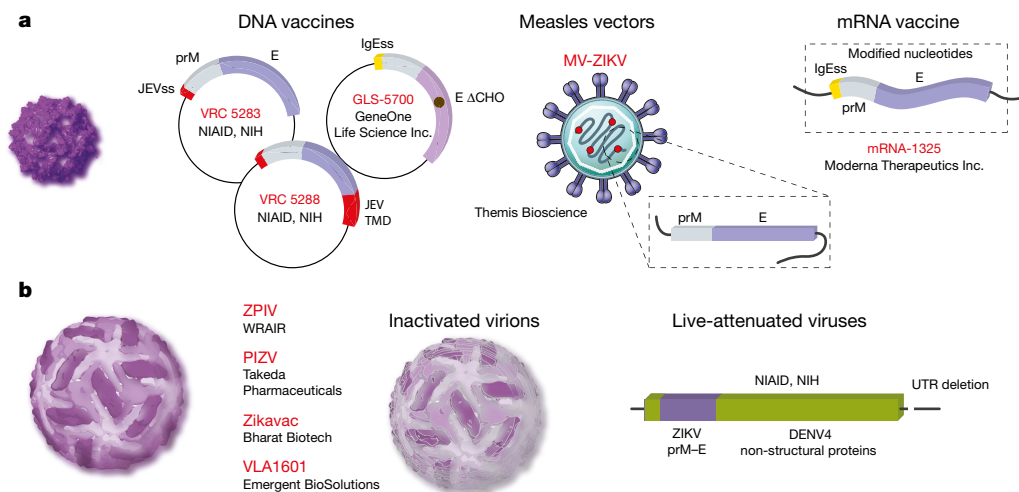


Fig. 4 | ZIKV vaccine platforms in clinical studies. **a**, Flavivirus prM–E proteins form non-infectious subviral particles that share functional and antigenic features with infectious virions. Subviral particles are smaller with $T = 1$ icosahedral symmetry, although they may be heterogeneous in size. Multiple ZIKV vaccine platforms that encode prM–E proteins have been evaluated in humans. DNA vaccines GLS-5700 (NCT02809443), VRC-5288 (NCT02840487) and VRC 5283 (NCT02996461) differ with respect to ZIKV strain and signal sequence preceding prM. The C terminus of VRC5288 is a chimaera of JEV. Nucleoside-modified mRNAs (mRNA-1325) and a measles vector (MV-ZIKV) expressing prM–E have

also been evaluated (NCT03014089 and NCT02996890, respectively).

b, Vaccine candidates derived from infectious ZIKV. Four inactivated vaccine candidates are being assessed. Phase I studies of the ZPIV vaccine construct developed by WRAIR have been conducted (NCT02963909, NCT02952833 and NCT02937233). Studies of the Takeda PIZV (NCT03343626), Emergent Biosolutions VLA1601 (NCT03425149) and Bharat Biotech ZikaVac are underway. Clinical trials of a chimeric live-attenuated vaccine derived from the NIAID DENV vaccine platform are anticipated.

infection in rhesus macaques⁷⁶, which has implications for ZIKV vaccine development and deployment. More detailed epidemiological evidence from humans is necessary to confirm whether clinically relevant ADE of ZIKV pathogenesis occurs, especially in the context of vertical transmission. To date, studies in Brazilian cohorts have not found any evidence of ADE, greater disease severity, or effects on birth outcomes in patients with acute ZIKV infection who had previously been exposed to DENV^{129,130}.

Development of ZIKV vaccines

Efforts to develop a vaccine were initiated rapidly after the threat of ZIKV congenital disease became clear. Multiple vaccine platforms have been evaluated in preclinical and clinical studies (Fig. 4). Because ZIKV circulates as a single serotype and infection provides immunity to re-challenge by heterologous strains, a requirement for only a single vaccine antigen is anticipated^{71,101}.

Synthetic nucleic acids provide a platform for the rapid development of vaccines. Multiple ZIKV prM–E DNA vaccine configurations have been evaluated that vary with respect to the sequence of the structural genes, method of codon optimization, signal sequence used to direct prM into the endoplasmic reticulum lumen and the plasmid backbone^{131,132}. A construct that lacks the ‘pr’ portion of prM has also been developed^{133,134}. Three of these DNA vaccines have been evaluated in human studies^{132,135}. DNA vaccine GLS-5700, which encodes a consensus of prM–E sequences from divergent ZIKV strains lacking the N-linked glycosylation site at E154, was safe and immunogenic in humans; neutralizing antibodies were present in 62% of recipients¹³². DNA vaccine candidates VRC5283 and VRC5288 encode a codon-optimized prM–E derived from an Asian strain downstream of the signal sequence of Japanese encephalitis virus (JEV). The stem and transmembrane domains at the C terminus of these constructs differ due to the replacement of this sequence in VRC5288 with the analogous sequence of JEV. Phase I studies revealed superior immunogenicity of VRC5283¹³⁵. Three doses of VRC5283 delivered at four-week intervals elicited neutralizing antibodies in all subjects. A phase II study to confirm immunogenicity and define efficacy is underway at twenty sites in the Americas.

Nucleoside-modified mRNAs drive protein expression at high levels in vivo, due in part to an ability to limit recognition by sensors

of foreign nucleic acids in transduced cells. Capped mRNAs are synthesized using modified nucleosides, encode a codon-optimized open reading frame flanked by untranslated regions, and are encapsulated in lipid nanoparticles or complexed with lipids. Multiple ZIKV mRNA vaccine candidates expressing prM–E elicit neutralizing antibodies at high titre and protect against viraemia after challenge^{136–138}. Similar to DNA vaccine platforms, these constructs differ with respect to the sequence of prM–E and signal sequence at the N terminus of prM; the contribution of these properties to differences in immunogenicity are not yet understood. Protection of NHPs following a single dose highlights the promise of this platform¹³⁶. One of these mRNAs (mRNA-1325) has been evaluated in phase I clinical studies.

The chemical inactivation of flaviviruses is an established method for creating protective immunogens for use in humans. Inactivated vaccines are currently in use against tick-borne encephalitis virus and JEV. The Walter Reed Army Institute of Research (WRAIR) developed an inactivated ZIKV vaccine by formalin-inactivation of a Puerto Rican ZIKV isolate. Administration of this ZIKV-purified inactivated vaccine (ZPIV) into mice and NHPs elicited neutralizing antibodies and conferred protection against viraemia following challenge^{133,134}. Subsequent studies demonstrated that two doses of ZPIV conferred protection for more than one year after vaccination¹³⁹. Safety and immunogenicity of ZPIV in humans was demonstrated in three placebo-controlled clinical studies of two doses at a 28-day interval. Although the WRAIR ZPIV candidate is not being evaluated further, similar inactivated vaccines are being developed by Takeda Pharmaceuticals (PIZV), Emergent Biosolutions (VLA1601) and Bharat Biotech. In contrast to other ZIKV vaccine candidates, the Bharat inactivated virus (ZikaVac) is derived from an African lineage strain.

Live-attenuated vaccines (LAVs) have been safe and cost-effective approaches to control flaviviruses. Whereas the first LAV was created by extensive passage of a strain of yellow fever virus in animals, molecular clone technology has enabled the rational design of attenuated vaccine candidates. ZIKV LAVs that use multiple attenuation strategies including deletions of the 3′ UTR, mutations to remove N-linked glycans on NS1, and chimerization with other flaviviruses have been evaluated in preclinical studies^{140–142}. In each case, vaccine-mediated

protection was established in mice or NHPs. Clinical studies of a ZIKV chimeric LAV (National Institute of Allergy and Infectious Diseases) will begin enrollment in 2018.

Viral vectored vaccines engineered to express ZIKV antigens also have promise. An adenovirus vector expressing ZIKV M–E elicited neutralizing antibodies and a protective immune response in mice and NHP models^{133,134}. Vesicular stomatitis and measles virus vectors expressing ZIKV prM–E are in preclinical development and phase I clinical trials, respectively¹⁴³. Vaccinia virus vectors expressing prM–E¹⁴⁴ or NS1¹⁴⁵ elicit protective responses in mouse models. Because NS1 is not present on virions, the NS1 antigen elicits antibodies that probably contribute to protection by recruiting host effector functions. Finally, partially purified recombinant proteins or virus-like particles elicit protective immune responses in mouse models¹⁴⁶.

A goal of ZIKV vaccine development is to protect against congenital disease. Most preclinical studies of ZIKV vaccine candidates evaluate protection from viraemia following peripheral challenge. The degree to which viral replication must be inhibited to prevent infection of the fetus or access to other immune privileged sites associated with persistence or transmission is unknown. The requirements for vaccine-mediated protection may also depend on the route of infection (mosquito versus sexual transmission). Several vaccines have the capacity to protect against vertical transmission of ZIKV to the fetus^{138,142}. Vaccine-elicited maternal immunity in the context of neonatal infection has also been shown using a vesicular stomatitis virus-vectored vaccine platform against ZIKV¹⁴³. Passive transfer of vaccine-elicited antibodies and the identification of neutralization titres that relate to protection suggest ZIKV-reactive antibody levels can be a functional correlate^{131,133,139}. It remains unclear whether estimates of a protective quantity of neutralizing antibody will apply uniformly to all vaccine platforms. Preclinical approaches to identify humoral and cellular correlates of protection may have a key role in vaccine licensure.

Conclusions

The epidemic of ZIKV and its clinical consequences resulted in a rapid research response, which has begun to provide answers as to why this virus transitioned from obscurity to notoriety. The scientific community is now answering questions related to viral evolution, structure and function, virulence, tropism and immune evasion, which begin to explain how ZIKV causes congenital disease. Unanswered questions remain with regard to transmission dynamics, viral persistence, cross-immunity with related viruses, as well as the neurodevelopmental sequelae of congenital infection. Although the last year has seen a waning of ZIKV cases, our new knowledge of ZIKV biology has informed the development of candidate vaccines and therapies, which will hopefully be implemented before a new epidemic. The lessons we have learned from ZIKV may be applicable to other viruses that cause future unanticipated clinical syndromes.

Received: 3 February 2018; Accepted: 19 July 2018;

Published online 29 August 2018.

- Weaver, S. C. et al. Zika virus: history, emergence, biology, and prospects for control. *Antiviral Res.* **130**, 69–80 (2016).
- Duffy, M. R. et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. *N. Engl. J. Med.* **360**, 2536–2543 (2009).
- Musso, D. et al. Zika virus in French Polynesia 2013–14: anatomy of a completed outbreak. *Lancet Infect. Dis.* **18**, e172–e182 (2018).
- Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
- Netto, E. M. et al. High Zika virus seroprevalence in Salvador, northeastern Brazil limits the potential for further outbreaks. *MBio* **8**, e01390–17 (2017).
- Annamalai, A. S. et al. Zika virus encoding non-glycosylated envelope protein is attenuated and defective in neuroinvasion. *J. Virol.* **e01348–17** (2017).
- Barba-Spaeth, G. et al. Structural basis of potent Zika–dengue virus antibody cross-neutralization. *Nature* **536**, 48–53 (2016).
- Fernandez, E. et al. Human antibodies to the dengue virus E-dimer epitope have therapeutic activity against Zika virus infection. *Nat. Immunol.* **18**, 1261–1269 (2017).
- Culshaw, A., Mongkolsapaya, J. & Screaton, G. R. The immunopathology of dengue and Zika virus infections. *Curr. Opin. Immunol.* **48**, 1–6 (2017).
- Faria, N. R. et al. Zika virus in the Americas: early epidemiological and genetic findings. *Science* **352**, 345–349 (2016).
- Prasad, V. M. et al. Structure of the immature Zika virus at 9 Å resolution. *Nat. Struct. Mol. Biol.* **24**, 184–186 (2017).
- Kostyuchenko, V. A. et al. Structure of the thermally stable Zika virus. *Nature* **533**, 425–428 (2016).
- Sirohi, D. et al. The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* **352**, 467–470 (2016).
- Two papers^{12,13} provide high-resolution cryo-EM structures of ZIKV.
- Rey, F. A., Stiasny, K. & Heinz, F. X. Flavivirus structural heterogeneity: implications for cell entry. *Curr. Opin. Virol.* **24**, 132–139 (2017).
- Aubry, M. et al. Zika virus seroprevalence, French Polynesia, 2014–2015. *Emerg. Infect. Dis.* **23**, 669–672 (2017).
- Swaminathan, S., Schlaberg, R., Lewis, J., Hanson, K. E. & Couturier, M. R. Fatal Zika virus infection with secondary nonsexual transmission. *N. Engl. J. Med.* **375**, 1907–1909 (2016).
- Carteaux, G. et al. Zika virus associated with meningoencephalitis. *N. Engl. J. Med.* **374**, 1595–1596 (2016).
- Karimi, O. et al. Thrombocytopenia and subcutaneous bleedings in a patient with Zika virus infection. *Lancet* **387**, 939–940 (2016).
- Dirlík, E. et al. Postmortem findings in patient with Guillain–Barré syndrome and Zika virus infection. *Emerg. Infect. Dis.* **24**, 114–117 (2018).
- Styczynski, A. R. et al. Increased rates of Guillain–Barré syndrome associated with Zika virus outbreak in the Salvador metropolitan area, Brazil. *PLoS Negl. Trop. Dis.* **11**, e0005869 (2017).
- Murray, K. O. et al. Prolonged detection of Zika virus in vaginal secretions and whole blood. *Emerg. Infect. Dis.* **23**, 99–101 (2017).
- Mansuy, J. M. et al. Zika Virus infection and prolonged viremia in whole-blood specimens. *Emerg. Infect. Dis.* **23**, 863–865 (2017).
- Michlmayr, D., Andrade, P., Gonzalez, K., Balmaseda, A. & Harris, E. CD14⁺CD16⁺ monocytes are the main target of Zika virus infection in peripheral blood mononuclear cells in a paediatric study in Nicaragua. *Nat. Microbiol.* **2**, 1462–1470 (2017).
- Miner, J. J. et al. Zika virus infection in mice causes panuveitis with shedding of virus in tears. *Cell Rep.* **16**, 3208–3218 (2016).
- Kodati, S. et al. Bilateral posterior uveitis associated with Zika virus infection. *Lancet* **389**, 125–126 (2017).
- Parke, D. W., III et al. Serologically confirmed Zika-related unilateral acute maculopathy in an adult. *Ophthalmology* **123**, 2432–2433 (2016).
- Tan, J. L. et al. Persistence of Zika virus in conjunctival fluid of convalescence patients. *Sci. Rep.* **7**, 11194 (2017).
- Mansuy, J. M. et al. Zika virus in semen and spermatozoa. *Lancet Infect. Dis.* **16**, 1106–1107 (2016).
- Mead, P. S. et al. Zika virus shedding in semen of symptomatic infected men. *N. Engl. J. Med.* **378**, 1377–1385 (2018).
- Hirsch, A. J. et al. Zika virus infection of rhesus macaques leads to viral persistence in multiple tissues. *PLoS Pathog.* **13**, e1006219 (2017).
- Govero, J. et al. Zika virus infection damages the testes in mice. *Nature* **540**, 438–442 (2016).
- Ma, W. et al. Zika virus causes testis damage and leads to male infertility in mice. *Cell* **167**, 1511–1524 (2016).
- Joguet, G. et al. Effect of acute Zika virus infection on sperm and virus clearance in body fluids: a prospective observational study. *Lancet Infect. Dis.* **17**, 1200–1208 (2017).
- Russell, K. et al. Male-to-female sexual transmission of Zika virus—United States, January–April 2016. *Clin. Infect. Dis.* **64**, 211–213 (2017).
- Deckard, D. T. et al. Male-to-male sexual transmission of Zika virus—Texas, January 2016. *MMWR Morb. Mortal. Wkly Rep.* **65**, 372–374 (2016).
- Oehler, E. et al. Zika virus infection complicated by Guillain–Barré syndrome—case report, French Polynesia, December 2013. *Euro Surveill.* **19**, 20720 (2014).
- Parra, B. et al. Guillain–Barré syndrome associated with Zika virus infection in Colombia. *N. Engl. J. Med.* **375**, 1513–1523 (2016).
- dos Santos, T. et al. Zika virus and the Guillain–Barré Syndrome — case series from seven countries. *N. Engl. J. Med.* **375**, 1598–1601 (2016).
- Description of ZIKV-associated Guillain–Barré Syndrome in the Americas.
- Dirlík, E. et al. Acute Zika virus infection as a risk factor for Guillain–Barré syndrome in Puerto Rico. *J. Am. Med. Assoc.* **318**, 1498–1500 (2017).
- Arora, N., Sadovsky, Y., Dermody, T. S. & Coyne, C. B. Microbial vertical transmission during human pregnancy. *Cell Host Microbe* **21**, 561–567 (2017).
- Miner, J. J. et al. Zika virus infection during pregnancy in mice causes placental damage and fetal demise. *Cell* **165**, 1081–1091 (2016).
- Establishment of a mouse model of the fetal injury caused by ZIKV.
- Sheridan, M. A. et al. Vulnerability of primitive human placental trophoblast to Zika virus. *Proc. Natl Acad. Sci. USA* **114**, E1587–E1596 (2017).
- Bayer, A. et al. Type III interferons produced by human placental trophoblasts confer protection against Zika virus infection. *Cell Host Microbe* **19**, 705–712 (2016).
- Jagger, B. W. et al. Gestational Stage and IFN- λ signaling regulate ZIKV infection in utero. *Cell Host Microbe* **22**, 366–376 (2017).
- Quicke, K. M. et al. Zika virus infects human placental macrophages. *Cell Host Microbe* **20**, 83–90 (2016).
- Richard, A. S. et al. AXL-dependent infection of human fetal endothelial cells distinguishes Zika virus from other pathogenic flaviviruses. *Proc. Natl Acad. Sci. USA* **114**, 2024–2029 (2017).

47. Martines, R. B. et al. Pathology of congenital Zika syndrome in Brazil: a case series. *Lancet* **388**, 898–904 (2016).
48. Platt, D. J. et al. Zika virus-related neurotropic flaviviruses infect human placental explants and cause fetal demise in mice. *Sci. Transl. Med.* **10**, ea07090 (2018).
49. Delaney, A. et al. Population-Based surveillance of birth defects potentially related to Zika virus infection — 15 States and U.S. Territories, 2016. *MMWR Morb. Mortal. Wkly Rep.* **67**, 91–96 (2018).
50. Li, H., Saucedo-Cuevas, L., Shrestha, S. & Gleeson, J. G. The neurobiology of Zika virus. *Neuron* **92**, 949–958 (2016).
51. Tang, H. et al. Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell* **18**, 587–590 (2016).
- Key paper describing ZIKV infection and injury of neuroprogenitor cells.**
52. Lum, F. M. et al. Zika virus infects human fetal brain microglia and induces inflammation. *Clin. Infect. Dis.* **64**, 914–920 (2017).
53. Meertens, L. et al. Axl mediates ZIKA virus entry in human glial cells and modulates innate immune responses. *Cell Rep.* **18**, 324–333 (2017).
54. Retallack, H. et al. Zika virus cell tropism in the developing human brain and inhibition by azithromycin. *Proc. Natl Acad. Sci. USA* **113**, 14408–14413 (2016).
55. Brasil, P. et al. Zika virus infection in pregnant women in Rio de Janeiro. *N. Engl. J. Med.* **375**, 2321–2334 (2016).
- Study describing the effects of ZIKV during pregnancy in Brazil.**
56. Cauchemez, S. et al. Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. *Lancet* **387**, 2125–2132 (2016).
57. Shapiro-Mendoza, C. K. et al. Pregnancy outcomes after maternal Zika virus infection during pregnancy — U.S. Territories, January 1, 2016–April 25, 2017. *MMWR Morb. Mortal. Wkly Rep.* **66**, 615–621 (2017).
58. Moura da Silva, A. A. et al. Early growth and neurologic outcomes of infants with probable congenital Zika virus syndrome. *Emerg. Infect. Dis.* **22**, 1953–1956 (2016).
59. Satterfield-Nash, A. et al. Health and development at age 19–24 months of 19 children who were born with microcephaly and laboratory evidence of congenital Zika virus infection during the 2015 Zika virus outbreak — Brazil, 2017. *MMWR Morb. Mortal. Wkly Rep.* **66**, 1347–1351 (2017).
60. Lazear, H. M. et al. A mouse model of Zika virus pathogenesis. *Cell Host Microbe* **19**, 720–730 (2016).
61. Honein, M. A. et al. Birth defects among fetuses and infants of US women with evidence of possible Zika virus infection during pregnancy. *J. Am. Med. Assoc.* **317**, 59–68 (2017).
62. Cugola, F. R. et al. The Brazilian Zika virus strain causes birth defects in experimental models. *Nature* **534**, 267–271 (2016).
- Establishment of a mouse model of fetal injury and microcephaly caused by ZIKV infection.**
63. Xavier-Neto, J. et al. Hydrocephalus and arthrogryposis in an immunocompetent mouse model of ZIKA teratogeny: a developmental study. *PLoS Negl. Trop. Dis.* **11**, e0005363 (2017).
64. Vermillion, M. S. et al. Intrauterine Zika virus infection of pregnant immunocompetent mice models transplacental transmission and adverse perinatal outcomes. *Nat. Commun.* **8**, 14575 (2017).
65. Szaba, F. M. et al. Zika virus infection in immunocompetent pregnant mice causes fetal damage and placental pathology in the absence of fetal infection. *PLoS Pathog.* **14**, e1006994 (2018).
66. Li, C. et al. Zika virus disrupts neural progenitor development and leads to microcephaly in mice. *Cell Stem Cell* **19**, 120–126 (2016).
67. Yockey, L. J. et al. Vaginal exposure to Zika virus during pregnancy leads to fetal brain infection. *Cell* **166**, 1247–1256 (2016).
- Animal study showing that intravaginal transmission of ZIKV can result in fetal brain injury.**
68. Gorman, M. J. et al. An immunocompetent mouse model of Zika virus infection. *Cell Host Microbe* **23**, 672–685 (2018).
69. Dudley, D. M. et al. A rhesus macaque model of Asian-lineage Zika virus infection. *Nat. Commun.* **7**, 12204 (2016).
70. Osuna, C. E. et al. Zika viral dynamics and shedding in rhesus and cynomolgus macaques. *Nat. Med.* **22**, 1448–1455 (2016).
71. Aliota, M. T. et al. Heterologous protection against Asian Zika virus challenge in rhesus macaques. *PLoS Negl. Trop. Dis.* **10**, e0005168 (2016).
72. Koide, F. et al. Development of a Zika virus infection model in cynomolgus macaques. *Front. Microbiol.* **7**, 2028 (2016).
73. Chiu, C. Y. et al. Experimental Zika virus inoculation in a new world monkey model reproduces key features of the human infection. *Sci. Rep.* **7**, 17126 (2017).
74. Li, X. F. et al. Characterization of a 2016 clinical isolate of Zika virus in non-human primates. *EBioMedicine* **12**, 170–177 (2016).
75. McCracken, M. K. et al. Impact of prior flavivirus immunity on Zika virus infection in rhesus macaques. *PLoS Pathog.* **13**, e1006487 (2017).
76. George, J. et al. Prior exposure to Zika virus significantly enhances peak dengue-2 viremia in rhesus macaques. *Sci. Rep.* **7**, 10498 (2017).
77. Driggers, R. W. et al. Zika virus infection with prolonged maternal viremia and fetal brain abnormalities. *N. Engl. J. Med.* **374**, 2142–2151 (2016).
78. Adams Waldorf, K. M. et al. Fetal brain lesions after subcutaneous inoculation of Zika virus in a pregnant nonhuman primate. *Nat. Med.* **22**, 1256–1259 (2016).
79. Nguyen, S. M. et al. Highly efficient maternal–fetal Zika virus transmission in pregnant rhesus macaques. *PLoS Pathog.* **13**, e1006378 (2017).
80. Martinot, A. J. et al. Fetal neuropathology in Zika virus-infected pregnant female rhesus monkeys. *Cell* **173**, 1111–1122 (2018).
81. Dudley, D. M. et al. Miscarriage and stillbirth following maternal Zika virus infection in nonhuman primates. *Nat. Med.* (2018).
82. Morrison, T. E. & Diamond, M. S. Animal models of Zika virus infection, pathogenesis, and immunity. *J. Virol.* **91**, e00009-17 (2017).
83. Abbink, P. et al. Therapeutic and protective efficacy of a dengue antibody against Zika infection in rhesus monkeys. *Nat. Med.* **24**, 721–723 (2018).
84. Mavigner, M. et al. Postnatal Zika virus infection is associated with persistent abnormalities in brain structure, function, and behavior in infant macaques. *Sci. Transl. Med.* **10**, ea06975 (2018).
85. Rossi, S. L. et al. Characterization of a novel murine model to study Zika virus. *Am. J. Trop. Med. Hyg.* **94**, 1362–1369 (2016).
86. Tripathi, S. et al. A novel Zika virus mouse model reveals strain specific differences in virus pathogenesis and host inflammatory immune responses. *PLoS Pathog.* **13**, e1006258 (2017).
87. Savidis, G. et al. The IFITMs inhibit Zika virus replication. *Cell Rep.* **15**, 2323–2330 (2016).
88. Monel, B. et al. Zika virus induces massive cytoplasmic vacuolization and paraptosis-like death in infected cells. *EMBO J.* **36**, 1653–1668 (2017).
89. Van der Hoek, K. H. et al. Viperin is an important host restriction factor in control of Zika virus infection. *Sci. Rep.* **7**, 4475 (2017).
90. Bowen, J. R. et al. Zika virus antagonizes type I interferon responses during infection of human dendritic cells. *PLoS Pathog.* **13**, e1006164 (2017).
91. Sun, X. et al. Transcriptional changes during naturally acquired Zika virus infection render dendritic cells highly conducive to viral replication. *Cell Rep.* **21**, 3471–3482 (2017).
92. Grant, A. et al. Zika virus targets human STAT2 to inhibit type I interferon signaling. *Cell Host Microbe* **19**, 882–890 (2016).
- A study that explains in part how ZIKV evades the interferon response in humans but not mice.**
93. Kumar, A. et al. Zika virus inhibits type-I interferon production and downstream signaling. *EMBO Rep.* **17**, 1766–1775 (2016).
94. Ding, Q. et al. Species-specific disruption of STING-dependent antiviral cellular defenses by the Zika virus NS2B3 protease. *Proc. Natl Acad. Sci. USA* **115**, E6310–E6318 (2018).
95. Xia, H. et al. An evolutionary NS1 mutation enhances Zika virus evasion of host interferon induction. *Nat. Commun.* **9**, 414 (2018).
96. Donald, C. L. et al. Full genome sequence and sRNA Interferon antagonist activity of Zika virus from Recife, Brazil. *PLoS Negl. Trop. Dis.* **10**, e0005048 (2016).
97. Zhu, Z. et al. Zika virus has oncolytic activity against glioblastoma stem cells. *J. Exp. Med.* **214**, 2843–2857 (2017).
98. Khan, S. et al. Dampened antiviral immunity to intravaginal exposure to RNA viral pathogens allows enhanced viral replication. *J. Exp. Med.* **213**, 2913–2929 (2016).
99. Rogers, T. F. et al. Zika virus activates de novo and cross-reactive memory B cell responses in dengue-experienced donors. *Sci. Immunol.* **2**, eaan6809 (2017).
100. Ricciardi, M. J. et al. Ontogeny of the B- and T-cell response in a primary Zika virus infection of a dengue-naïve individual during the 2016 outbreak in Miami, FL. *PLoS Negl. Trop. Dis.* **11**, e0006000 (2017).
101. Dowd, K. A. et al. Broadly neutralizing activity of Zika virus–immune sera identifies a single viral serotype. *Cell Rep.* **16**, 1485–1491 (2016).
102. Priyamvada, L., Suthar, M. S., Ahmed, R. & Wrammert, J. Humoral immune responses against Zika virus infection and the importance of preexisting flavivirus immunity. *J. Infect. Dis.* **216**, S906–S911 (2017).
103. Stettler, K. et al. Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. *Science* **353**, 823–826 (2016).
104. Sappapapu, G. et al. Neutralizing human antibodies prevent Zika virus replication and fetal disease in mice. *Nature* **540**, 443–447 (2016).
- First two papers^{103,104} describing neutralizing human monoclonal antibodies against ZIKV.**
105. Lai, L. et al. Innate, T-, and B-cell responses in acute human Zika patients. *Clin. Infect. Dis.* **66**, 1–10 (2018).
106. Priyamvada, L. et al. Human antibody responses after dengue virus infection are highly cross-reactive to Zika virus. *Proc. Natl Acad. Sci. USA* **113**, 7852–7857 (2016).
107. Zhao, H. et al. Structural basis of Zika virus-specific antibody protection. *Cell* **166**, 1016–1027 (2016).
108. Swanson, J. A. et al. Dengue Virus envelope dimer epitope monoclonal antibodies isolated from dengue patients are protective against Zika virus. *MBio* **7**, e01123-16 (2016).
109. Collins, M. H. et al. Lack of durable cross-neutralizing antibodies against Zika virus from dengue virus infection. *Emerg. Infect. Dis.* **23**, 773–781 (2017).
110. Rouvinski, A. et al. Recognition determinants of broadly neutralizing human antibodies against dengue viruses. *Nature* **520**, 109–113 (2015).
111. Wang, J. et al. A Human bi-specific antibody against Zika virus with high therapeutic potential. *Cell* **171**, 229–241 (2017).
112. Yu, L. et al. Delineating antibody recognition against Zika virus during natural infection. *JCI Insight* **2**, 93042 (2017).
113. Wang, Q. et al. Molecular determinants of human neutralizing antibodies isolated from a patient infected with Zika virus. *Sci. Transl. Med.* **8**, 369ra179 (2016).
114. Pardy, R. D. et al. Analysis of the T cell response to Zika virus and identification of a novel CD8⁺ T cell epitope in immunocompetent mice. *PLoS Pathog.* **13**, e1006184 (2017).
115. Elong Ngono, A. et al. Mapping and role of the CD8⁺ T cell response during primary Zika virus infection in mice. *Cell Host Microbe* **21**, 35–46 (2017).

116. Huang, H. et al. CD8⁺ T cell immune response in immunocompetent mice during Zika virus infection. *J. Virol.* **91**, e00900-17 (2017).
117. Wen, J. et al. Identification of Zika virus epitopes reveals immunodominant and protective roles for dengue virus cross-reactive CD8⁺ T cells. *Nat. Microbiol.* **2**, 17036 (2017).
118. Manangeeswaran, M., Ireland, D. D. & Verthelyi, D. Zika (PRVABC59) infection is associated with T cell infiltration and neurodegeneration in CNS of immunocompetent neonatal C57BL/6 mice. *PLoS Pathog.* **12**, e1006004 (2016).
119. Jurado, K. A. et al. Antiviral CD8 T cells induce Zika-virus-associated paralysis in mice. *Nat. Microbiol.* **3**, 141–147 (2018).
120. Cimini, E. et al. Human Zika infection induces a reduction of IFN- γ producing CD4 T-cells and a parallel expansion of effector V α 2 T-cells. *Sci. Rep.* **7**, 6313 (2017).
121. Grifoni, A. et al. Prior Dengue virus exposure shapes T cell immunity to Zika virus in humans. *J. Virol.* e01469-17 (2017).
122. Weiskopf, D. et al. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8⁺ T cells. *Proc. Natl Acad. Sci. USA* **110**, E2046–E2053 (2013).
123. Liu, Y. et al. Evolutionary enhancement of Zika virus infectivity in *Aedes aegypti* mosquitoes. *Nature* **545**, 482–486 (2017).
124. Yuan, L. et al. A single mutation in the prM protein of Zika virus contributes to fetal microcephaly. *Science* **358**, 933–936 (2017).
Two papers^{123,124} describe the genetic changes in epidemic ZIKV strains that may explain altered epidemiology and pathogenicity.
125. Klase, Z. A. et al. Zika fetal neuropathogenesis: etiology of a viral syndrome. *PLoS Negl. Trop. Dis.* **10**, e0004877 (2016).
126. Chavali, P. L. et al. Neurodevelopmental protein Musashi-1 interacts with the Zika genome and promotes viral replication. *Science* **357**, 83–88 (2017).
127. Dejnirattisai, W. et al. Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with Zika virus. *Nat. Immunol.* **17**, 1102–1108 (2016).
128. Bardina, S. V. et al. Enhancement of Zika virus pathogenesis by preexisting ant flavivirus immunity. *Science* **356**, 175–180 (2017).
129. Terzian, A. C. B. et al. Viral load and cytokine response profile does not support antibody-dependent enhancement in dengue-primed Zika virus-infected patients. *Clin. Infect. Dis.* **65**, 1260–1265 (2017).
130. Halai, U. A. et al. Maternal Zika virus disease severity, virus load, prior dengue antibodies, and their relationship to birth outcomes. *Clin. Infect. Dis.* **65**, 877–883 (2017).
131. Dowd, K. A. et al. Rapid development of a DNA vaccine for Zika virus. *Science* **354**, 237–240 (2016).
132. Tebas, P. et al. Safety and immunogenicity of an anti-Zika virus DNA vaccine — preliminary report. *N. Engl. J. Med.* <https://doi.org/10.1056/NEJMoa1708120> (2017).
133. Abbink, P. et al. Protective efficacy of multiple vaccine platforms against Zika virus challenge in rhesus monkeys. *Science* **353**, 1129–1132 (2016).
134. Larocca, R. A. et al. Vaccine protection against Zika virus from Brazil. *Nature* **536**, 474–478 (2016).
135. Gaudinski, M. R. et al. Safety, tolerability, and immunogenicity of two Zika virus DNA vaccine candidates in healthy adults: randomised, open-label, phase 1 clinical trials. *Lancet* **391**, 552–562 (2018).
Five papers^{131–135} describe the DNA and inactivated vaccine platforms under development against ZIKV.
136. Pardi, N. et al. Zika virus protection by a single low-dose nucleoside-modified mRNA vaccination. *Nature* **543**, 248–251 (2017).
137. Richner, J. M. et al. Modified mRNA vaccines protect against Zika virus infection. *Cell* **168**, 1114–1125 (2017).
138. Richner, J. M. et al. Vaccine mediated protection against Zika virus-induced congenital disease. *Cell* **170**, 273–283 (2017).
Three papers^{136–138} describe the use of mRNA-based vaccines against ZIKV.
139. Abbink, P. et al. Durability and correlates of vaccine protection against Zika virus in rhesus monkeys. *Sci. Transl. Med.* **9**, eaao4163 (2017).
140. Xie, X. et al. Understanding Zika virus stability and developing a chimeric vaccine through functional analysis. *MBio* **8**, e02134-16 (2017).
141. Shan, C. et al. A live-attenuated Zika virus vaccine candidate induces sterilizing immunity in mouse models. *Nat. Med.* **23**, 763–767 (2017).
142. Shan, C. et al. A single-dose live-attenuated vaccine prevents Zika virus pregnancy transmission and testis damage. *Nat. Commun.* **8**, 676 (2017).
143. Betancourt, D., de Queiroz, N. M., Xia, T., Ahn, J. & Barber, G. N. Cutting edge: innate immune augmenting vesicular stomatitis virus expressing Zika virus proteins confers protective immunity. *J. Immunol.* **198**, 3023–3028 (2017).
144. Prow, N. A. et al. A vaccinia-based single vector construct multi-pathogen vaccine protects against both Zika and chikungunya viruses. *Nat. Commun.* **9**, 1230 (2018).
145. Brault, A. C. et al. A Zika vaccine targeting NS1 protein protects immunocompetent adult mice in a lethal challenge model. *Sci. Rep.* **7**, 14769 (2017).
146. Salvo, M. A., Kingstad-Bakke, B., Salas-Quinchucua, C., Camacho, E. & Osorio, J. E. Zika virus like particles elicit protective antibodies in mice. *PLoS Negl. Trop. Dis.* **12**, e0006210 (2018).
147. Bayer, A. et al. Chromosome 19 microRNAs exert antiviral activity independent from type III interferon signaling. *Placenta* **61**, 33–38 (2018).

Acknowledgements This work was supported by NIH grants (R01 AI073755, R01 AI104972, U19 AI083019 and R01 HD091218) and by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH. We thank E. Tyler (NIH) for assistance with figure preparation of virion models. This publication is the responsibility of the authors and does not necessarily represent the official view of the NIH.

Reviewer information Nature thanks J. Jung and H. Tang for their contribution to the peer review of this work.

Author contributions T.C.P. and M.S.D. conceived and wrote the review.

Competing interests M.S.D. is a consultant for Inbios and on the Scientific Advisory Board of Moderna. T.C.P. is a co-inventor of NIAID ZIKV vaccine candidates.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to T.C.P. and M.S.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Measurements of the gravitational constant using two independent methods

Qing Li^{1,8}, Chao Xue^{2,3,8}, Jian-Ping Liu^{1,8}, Jun-Fei Wu^{1,8}, Shan-Qing Yang^{1*}, Cheng-Gang Shao^{1*}, Li-Di Quan⁴, Wen-Hai Tan¹, Liang-Cheng Tu^{1,2}, Qi Liu^{2,3}, Hao Xu¹, Lin-Xia Liu⁵, Qing-Lan Wang⁶, Zhong-Kun Hu¹, Ze-Bing Zhou¹, Peng-Shun Luo¹, Shu-Chao Wu¹, Vadim Milyukov⁷ & Jun Luo^{1,2,3*}

The Newtonian gravitational constant, G , is one of the most fundamental constants of nature, but we still do not have an accurate value for it. Despite two centuries of experimental effort, the value of G remains the least precisely known of the fundamental constants. A discrepancy of up to 0.05 per cent in recent determinations of G suggests that there may be undiscovered systematic errors in the various existing methods. One way to resolve this issue is to measure G using a number of methods that are unlikely to involve the same systematic effects. Here we report two independent determinations of G using torsion pendulum experiments with the time-of-swing method and the angular-acceleration-feedback method. We obtain G values of 6.674184×10^{-11} and 6.674484×10^{-11} cubic metres per kilogram per second squared, with relative standard uncertainties of 11.64 and 11.61 parts per million, respectively. These values have the smallest uncertainties reported until now, and both agree with the latest recommended value within two standard deviations.

A precise knowledge of G is not only of considerable metrological interest, but also important because of the key role of G in fields such as gravitation, cosmology, particle physics, geophysics and astrophysics. However, this constant is difficult to measure accurately because of the extreme weakness and non-shieldability of gravity. The first G value, with an uncertainty of about 1%, was obtained from Cavendish and Michell's torsion pendulum experiment¹ in 1798. Since then, more than 200 experiments have been performed to determine G ^{2,3}. However, the uncertainty of G has been reduced by a factor of only about 10 per century. In 2016, the Committee on Data for Science and Technology published an updated G value (CODATA-2014) of $6.67408(31) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ with a relative uncertainty of 47 parts per million (p.p.m.)⁴, which is still many orders of magnitude larger than that of other important fundamental constants.

In the CODATA-2014 adjustment, fourteen values of G determined in the past four decades are considered with smallest relative uncertainty of 14 p.p.m. However, the difference between the largest and the smallest G values is close to 550 p.p.m., which is almost 40 times the magnitude of the smallest uncertainty. Up to now, no well established physical theory or mechanism has been able to explain such a wide-range scattering of the G value. The most probable explanation lies in undiscovered systematic errors in all or some of these experiments. In view of the different error sources in different experiments, the only way to solve this problem and improve the confidence level, as discussed by Quinn et al.^{5–8}, is to measure the constant using a number of different methods. At the International Bureau of Weights and Measures, Quinn and colleagues have measured G with two methods^{9–11} and obtained results at the high end of the G values adopted in the CODATA-2014 adjustment. In this work, we performed a new determination of G

using torsion pendulum experiments on different apparatus with two completely independent methods (see Supplementary Information Section 1 and Supplementary Tables 1–3)—the time-of-swing (TOS) method and the angular-acceleration-feedback (AAF) method—so that unknown systematic errors in one method would be unlikely to exist in the other.

The TOS method, most famously used by Heyl^{12,13} in the 1930s, measures the change in the torsional oscillation frequency of a pendulum with the source masses arranged in two different configurations: the ‘near’ position, where the source masses are in line with the equilibrium position of the torsion pendulum, leading to a faster oscillation, and the ‘far’ position, where the source masses are perpendicular to the equilibrium position of the torsion pendulum, resulting in a slower oscillation. The AAF method was first used to measure G by Rose et al.¹⁴ in 1969 and was considerably improved by Gundlach et al.¹⁵ In this method, two turntables are used to rotate the torsion pendulum coaxially and the source masses individually. With a high-gain feedback control system, the twist angle of the fibre is reduced to about zero and thus the angular acceleration of the pendulum is equal to the gravitational angular acceleration generated by the source masses.

Experimental challenge and solution

Since the 1980s, our group has been measuring G with the TOS method and has obtained many phased results^{16–20}. To reduce the anelastic effect (the frequency-dependent property of the torsion spring constant)^{21–23} of the fibre, fused silicon dioxide (silica) fibres with high quality factor of the torsional oscillation mode (Q) were used in the present measurements, which were performed on two independent apparatus (Extended Data Fig. 1a–c). In the experiment using apparatus 1

¹MOE Key Laboratory of Fundamental Physical Quantities Measurements, Hubei Key Laboratory of Gravitation and Quantum Physics, School of Physics, Huazhong University of Science and Technology, Wuhan, China. ²TianQin Research Center for Gravitational Physics, Sun Yat-sen University, Zhuhai, China. ³School of Physics and Astronomy, Sun Yat-sen University, Zhuhai, China. ⁴College of Engineering, Huzhou University, Huzhou, China. ⁵Teaching Research and Assessment Center, Henan Institute of Technology, Xinxiang, China. ⁶School of Science, Hubei University of Automotive Technology, Shiyan, China. ⁷Sternberg Astronomical Institute, Moscow State University, Moscow, Russia. ⁸These authors contributed equally: Qing Li, Chao Xue, Jian-Ping Liu, Jun-Fei Wu. *e-mail: ysq2011@hust.edu.cn; cgshao@hust.edu.cn; junluo@hust.edu.cn

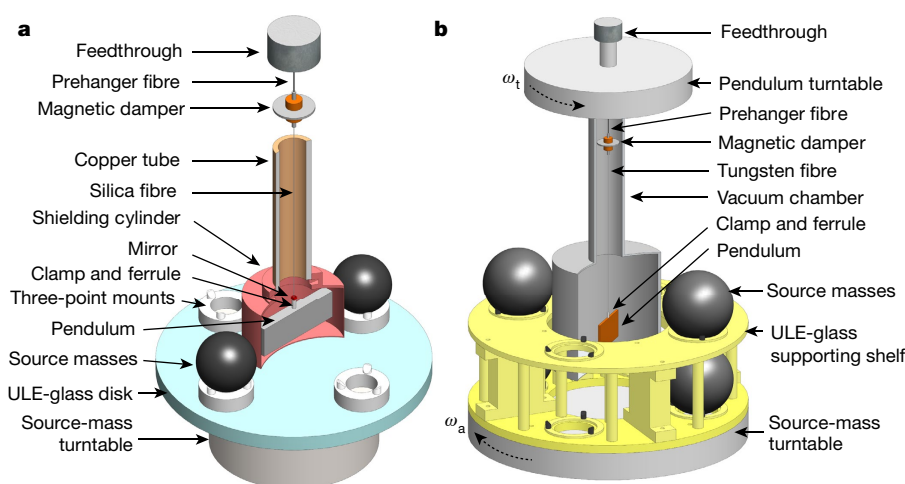


Fig. 1 | Sketch of the experiment. **a**, In the TOS method, the pendulum is an Al-coated fused silica block with dimensions of $91 \times 11 \times 31 \text{ mm}^3$ and mass of about 68 g. The pendulum is suspended by a thin fused silica fibre with a diameter of 40–60 μm and a length of 900 mm. The magnetic damper is suspended through a 50-mm-long, 80- μm -diameter tungsten fibre. Two SS316 stainless-steel spheres with an average diameter of 57.2 mm and a vacuum mass of 778 g are used as the source masses. A turntable is used to change the positions of the spheres between the ‘near’ and ‘far’ configurations (the ‘near’ configuration is shown here; in the ‘far’ configuration, the turntable is rotated by 90°). A hollow gold-coated aluminium cylinder installed between the pendulum and the spheres is used to shield the system from the electrostatic field. The pendulum and the source masses are placed inside the same vacuum

chamber with a pressure of about 10^{-5} Pa maintained by an ion pump. The pendulum twist is monitored by an optical lever. **b**, In the AAF method, the pendulum is a gold-coated fused silica block with dimensions of $91 \times 4 \times 50 \text{ mm}^3$ and a vacuum mass of 40 g. The main fibre is an 870-mm-long, 25- μm -diameter tungsten fibre. The design of the magnetic damper is the same as that in the TOS method. Four SS316 stainless-steel spheres with an average diameter of 127.0 mm and a vacuum mass of 8,541 g are used as the source masses that sit on an ULE-material shelf with upper and lower layers. The small deflection angle of the pendulum is recorded by an autocollimator. The chamber with the pendulum is hung under an air-bearing turntable, which is installed coaxially with the separate source-mass turntable. The apparatus are located in the passive thermal room situated in our cave laboratory.

(TOS-I), three different silica fibres were used to check for possible fibre-induced errors while all other parts of the apparatus were unchanged for all measurements. Apparatus 2 was placed in another room, about 150 m away from apparatus 1. In the experiment with apparatus 2 (TOS-II), a new silica fibre with another set of pendulum and source masses was used to test for possible errors dependent on the apparatus. Furthermore, we minimized other large systematic uncertainties encountered in our previous experiment^{18,19}.

Since 2008, our group has been conducting proof-of-principle experiments with the AAF method^{24,25}. In this work, the apparatus was redesigned and completely rebuilt (Extended Data Fig. 1d–f) to reduce several sources of uncertainty that existed in our previous measurements: (1) the aluminium shelf supporting the source masses was substituted with an ultra-low thermal expansion (ULE) glass shelf to reduce the influence of temperature on the distance between the source masses; (2) the turntable supporting the vacuum chamber and the pendulum was replaced by a large hollow-bowl air bearing and moved from the bottom to the top of the apparatus to improve stability; (3) two different methods were used to measure the distance between the source masses and thus improve the confidence level; and (4) the co-moving background gravity gradient created by the rotating shelf was compensated directly to reduce its effect on the G measurement. With the AAF method, we measured the G value at three different conditions (referred to as AAF-I, AAF-II and AAF-III). The selected signal frequency in AAF-I was different from the other two measurements. In AAF-III, other members of the group repeated the measurement of G with two additional improvements: the magnetic damper correction was reduced by optimizing the prehanger fibre and the magnetic effect was reduced by adding a Mu-metal shield around the pendulum.

Schematics of the two methods are shown in Fig. 1. In both methods, the heart of the apparatus is a two-stage pendulum system that consists of a magnetic damper and a torsion pendulum. The passive magnetic damper is used to suppress the swinging mode of the torsion pendulum, which is excited by ambient vibration noise²⁶. Well characterized stainless-steel spheres are used as the source masses. Because the determination of G is based on Newton’s formula, $F = G M m / r^2$ (where F is

the gravitational force between masses M and m , which are located at a distance r), we need to measure the dimensions, density, homogeneity and relative positions of the spheres with sufficient accuracy. For this purpose, considerable efforts were devoted to grinding and polishing the pendulum block and the source masses to obtain a perfect geometry (Extended Data Tables 1, 2). The assembly and alignment of the pendulum and source masses were carried out with great care, following the method introduced in ref.¹⁹ (Supplementary Information Section 2). To eliminate possible human errors, almost all parameters were measured repeatedly by different members of the group, and the combined uncertainties are shown in Table 1.

The silica fibre, a critical component in the TOS method, was pulled from a high-purity fused silica rod using an oxygen–natural gas flame (Extended Data Fig. 2). Four fibres with diameters of 40–60 μm , lengths of 900 mm and $Q = (2–3) \times 10^5$ were selected for the experiments to obtain an optimal signal-to-noise ratio. The fibre surfaces were sputter-coated by 5-nm-thick germanium and then 10-nm-thick bismuth to suppress the electrostatic influence from the charges accumulated on the surfaces of the pendulum and fibre. The Ge buffer layer kept the interface dissipation low, and the conductive Bi layer enabled charge flow²⁷. After coating, the quality factors were decreased to $(3–6) \times 10^4$, but they were still one order of magnitude higher than that of the tungsten fibre used in our previous experiment^{18,19} ($Q \approx 1.7 \times 10^3$). Considering the correction factor $1/(\pi Q)$ proposed by Kuroda²¹, we estimate the correction to the G value due to anelasticity to be 5–9 p.p.m., and half of this value is treated as the uncertainty (Extended Data Table 3).

In the AAF method, precision control of the turntable rotation is a key factor. It is realized by using two feedback loops with the proportion–integration–differentiation control algorithm²⁵. The angular velocity, $\omega_t(t)$, of the pendulum turntable is feedback-controlled to minimize the twist angle of the fibre to about zero (Fig. 2d). Meanwhile, the angular velocity, $\omega_a(t)$, of the source-mass turntable is controlled to maintain a constant difference (ω_d) between the angular velocities of the two turntables so that $\omega_a(t) = \omega_d + \omega_t(t)$. Both $\omega_t(t)$ and $\omega_a(t)$ vary sinusoidally and have the same amplitude. When the two feedback

Table 1 | Contributions of various experimental parameters to the main error budget of the measurements, expressed in parts per million

Parameter	TOS-I Fibre 1	TOS-I Fibre 2	TOS-I Fibre 3	TOS-II Fibre 4	AAF-I	AAF-II	AAF-III
Pendulum							
Dimensions	1.82	1.82	1.82	2.73	0.16	0.16	0.16
Attitude	0.01 [0.02]	0.01	0.05 [0.03]	0.02	0.06	0.06	0.03
Density inhomogeneity	0.20	0.20	0.20	0.20	0.46	0.46	0.46
Coating layer	0.86	0.86	0.86	0.73	0.34	0.34	0.34
Clamp and ferrule	0.15	0.15	0.15	0.33	0.70	1.05	0.48
Others	0.40	0.37	0.39	0.26	0.29	0.29	0.29
Source masses							
Masses	0.73	0.73	0.55	0.55	0.32	0.31	0.31
Horizontal distance	8.73	8.73	8.47	9.53 [9.27]	8.98	8.98	8.98
Vertical distance	—	—	—	—	5.79	5.79	5.79
Positions, alignment	1.51 [1.60]	0.64	1.81 [1.85]	0.63 [0.68]	0.57	0.62	0.35
Fibre nonlinearity	1.45	4.84	1.10 [1.03]	1.67 [1.26]	—	—	—
Fibre anelasticity	3.00	4.19	2.84	3.46	0.01	0.01	0.01
Thermal effect	0.71	3.41	0.77 [0.61]	0.97 [1.46]	0.91	0.91	0.91
Time base	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Gravitational nonlinearity	0.62	0.62	0.62	0.22	—	—	—
Rotating gravity gradient	—	—	—	—	1.86	1.35	1.72
Shelf deformation	—	—	—	—	1.51	1.51	1.51
Magnetic damper	0.08	1.19	0.05	0.08	1.95	1.95	0.08
Air density	—	—	—	—	1.00	1.51	1.13
Magnetic field	2.08	2.08	2.08	0.71	3.98	3.98	0.90
Electrostatic field	0.17	0.17	0.17	0.17	—	—	—
Angle encoder	—	—	—	—	0.72	0.72	0.72
Residual twist angle	—	—	—	—	0.03	0.61	0.45
Statistical error of $\Delta\omega^2$ or α_t	10.22 [10.83]	30.67	12.03 [10.22]	13.78 [13.78]	3.44	2.60	1.34
Total	14.29 [14.74]	32.88	15.46 [14.09]	17.49 [17.35]	12.45	12.27	11.21
Combined uncertainty	13.67	32.88	13.96	15.59	—	—	—

For fibres 1, 3 and 4, each G measurement was performed twice with random orientations of the source masses. The values in the square brackets represent the values obtained in the repeated experiments. Uncertainties are one standard deviation. 'Others' includes effects due to the pendulum mass, the reflecting mirror, glues, edge flaws and the silica rod. 'Thermal effect' includes the fibre thermoelasticity in the TOS method. In the AAF method, the fibre thermoelasticity is negligible because the fibre does not twist, and the thermal effect is evaluated by modulating the temperature in the room.

loops work well cooperatively, the angular acceleration signal of the pendulum turntable of interest appears at $2\omega_d$ with an amplitude of about 462 nrad s^{-2} , and is quantified by the gravitational interaction strength between the pendulum and the spheres. This scheme helps to clearly separate the signal from the laboratory-fixed gravitational background and other similar noises in the frequency domain.

In this experiment, ω_d was usually set to a few milliradians per second so that the signal frequency ($2\omega_d$) was in a frequency (f) range with low $1/f$ noise inherent in the torsion fibre. For most of the experimental runs, $\omega_d = 5.235988(4) \text{ mrad s}^{-1}$ and the signal frequency was about 1.67 mHz (uncertainties are 1σ unless stated otherwise). The average values of $\omega_t(t)$ and $\omega_a(t)$ were about 2.44 mrad s^{-1} and $-2.79 \text{ mrad s}^{-1}$ (the minus sign denotes opposite rotation direction), respectively, which were chosen to be far from the harmonic signals of the laboratory-fixed background and make the turntable operate at appropriate rotating speeds. Furthermore, when we used an angular velocity difference of $\omega_d = 7.853982(3) \text{ mrad s}^{-1}$ (where $\omega_t(t) \approx 3.49 \text{ mrad s}^{-1}$ and $\omega_a(t) \approx -4.36 \text{ mrad s}^{-1}$), the signal frequency was about 2.50 mHz in AAF-I, and we found no dependence of the result on angular velocity.

In both methods, the relative position of the spheres to the pendulum is much less critical, but the distance between the geometric centres of the spheres (Extended Data Fig. 3) must be measured with high accuracy. To improve the position stability of the spheres, updated three-point mounts were used to support the spheres. The position repeatability and the influence of temperature and vibration were investigated in detail²⁸. Furthermore, a ULE-material disk or shelf was used to support the three-point mounts to reduce the temperature influence on the distance. In the TOS method, the distance of the geometric

centres of the spheres was measured before and after each experiment by using the rotating gauge block method²⁹ with an uncertainty of less than $0.4 \mu\text{m}$. In the AAF method, four distances (Extended Data Fig. 3) between the geometric centres of the four spheres were determined using a coordinate measuring machine with an uncertainty of less than $2.0 \mu\text{m}$. The horizontal separations were verified with the rotating gauge block method, and the vertical surface separations were checked by inserting a small gauge block ($1\text{--}2 \mu\text{m}$ thinner than the gap) in the gap between the sphere surfaces. The results obtained with different methods agree with each other within $2 \mu\text{m}$. The temperature effect on the distances of the geometric centres of the spheres was investigated by temperature modulation experiments (Extended Data Table 4).

Main systematic errors

The analysis of systematic effects is crucial in measuring the intrinsically weak gravitational force. Some of the main systematic errors are discussed in the following (Extended Data Table 3).

Density inhomogeneity of the pendulum and source masses

In both methods, the density inhomogeneity of the pendulum body and the source masses influences the accuracy in calculating the gravitational torque and the moment of inertia of the pendulum. The planar density distribution of the glass pendulum was measured by the optical interference method³⁰, which provides an uncertainty of less than 0.5 p.p.m. in both methods. The density inhomogeneity of the source masses was measured using three methods: (i) scanning slices cut from the sample sphere with scanning electron microscopy³¹, which yields an uncertainty of less than 0.1 p.p.m. to the value of G ; (ii) measuring

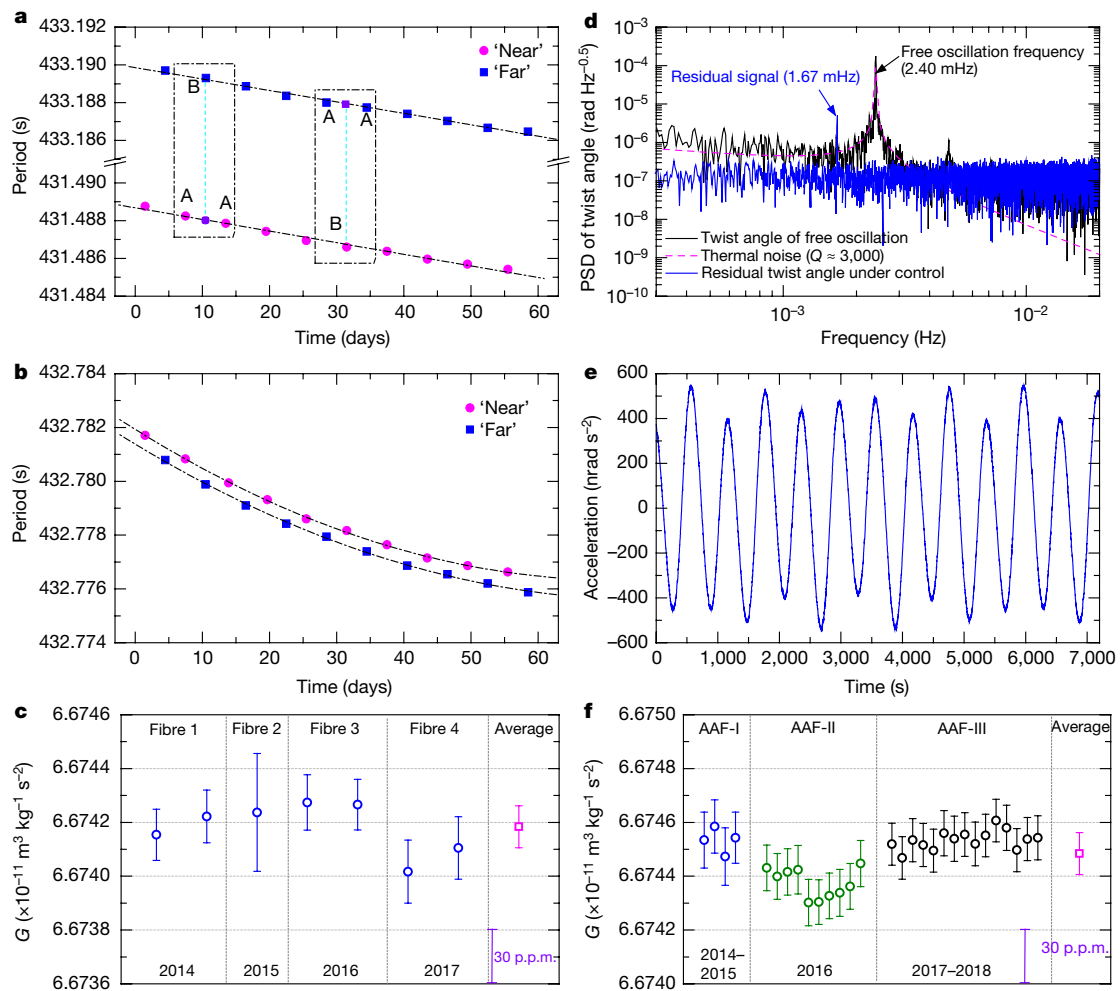


Fig. 2 | Experimental data. **a**, Typical periods extracted from 10 sets of time-series angle data in the TOS method for one fibre. The period difference between the 'near' and 'far' positions is about 1.7 s. The statistical uncertainty of each data point extracted from a three-day data segment is about 0.03 ms. The 'A-B-A' method¹⁹ is used to determine the period difference and reduce the effect of the period drift (dot-dashed lines) due to the 'aging' effect of the fibre. **b**, Typical sets of background periods measured without the source masses. **c**, The 7 values of G obtained using four fibres and the TOS method. The measurement was carried out once with fibre 2 and twice with random orientations of the source masses for fibres 1, 3 and 4. **d**, The typical power spectral density (PSD) of the

twist angles of the pendulum for the AAF method. At the signal frequency of interest, the typical residual twist angle of the pendulum is 17.1(3) nrad, contributing a correction of 4.37(9) p.p.m. to the value of G . **e**, Two-hour segment of the angular acceleration data of the torsion pendulum turntable. The curve is jagged mainly owing to the mixture of the laboratory-fixed environment gravitational gradient signal with the signal of interest (Supplementary Information Section 5, Supplementary Fig. 2). **f**, The values of G obtained by the AAF method. Each point denotes the value of G obtained with different orientations of the spheres. The signal frequency is about 2.50 MHz in AAF-I and ~1.67 MHz in AAF-II and AAF-III. All error bars denote 1σ confidence level.

the offset of the centre of mass from the geometric centre by using a beam balance³²; (iii) measuring the same centre offset by using the air-bearing method. The eccentricities of the source masses determined by methods (ii) and (iii) are less than 0.3 μm in the TOS method and less than 1.3 μm in the AAF method. These eccentricities are mainly caused by nonsphericities, which were considered in the determination of the geometric centre distance between the spheres. Furthermore, the orientations of the spheres were changed randomly before each run to further average out the effects of density inhomogeneity and nonsphericity.

Magnetic damper

The magnetic damper, which is generally used to suppress the swinging modes of the torsion pendulum, introduces an additional effect to the G measurement. The correction for this effect is $I_m K^2 / (I K_m^2)$ in the TOS method¹⁹ and $I_m K / (I K_m)$ in the AAF method²⁴, where I and I_m are the moments of inertia of the pendulum and the magnetic damper and K and K_m are the torsion spring constants of the main fibre and the prehanger fibre, respectively. In the TOS method, we choose a ~50-mm-long, 80- μm -diameter tungsten fibre as the prehanger fibre,

and a correction of only a few parts per million is required to the G value. In the AAF method, we use the same design for the magnetic damper and the prehanger fibre, which contributes a correction of 455.40(1.95) p.p.m. to the G value in AAF-I and AAF-II. This correction is reduced to 25.74(8) p.p.m. by decreasing the length (~35 mm) and increasing the diameter (150 μm) of the prehanger fibre in AAF-III (Supplementary Table 1).

Coating layer on the pendulum

The surface of the pendulum is coated with a thin metal film to eliminate the electrostatic effect. Gold is commonly used as a coating material to achieve a smooth conductive surface. The coating layer increases the moment of inertia of the pendulum and the gravitational torque exerted by the source masses. In the AAF method, a ~400-nm-thick Au/Cu layer (Cu is the sublayer) is coated on the pendulum surface, which introduces a correction of -9.10(34) p.p.m., as evaluated according to the thickness distribution and the mass of the coating layer. In a previous experiment³³ using the TOS method, the Au/Cu coating layer introduced a correction of -24.28(4.33) p.p.m. to the G value. In this work, a ~200-nm-thick aluminium layer is used to replace the Au/Cu

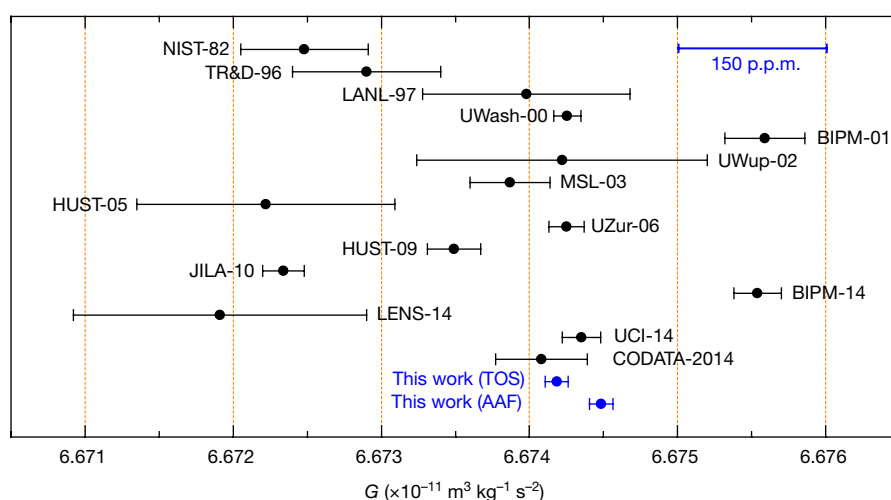


Fig. 3 | Comparison with previous results. G values obtained in this work compared with recent measurements (NIST-82³⁹, TR&D-96⁴⁰, LANL-97⁴¹, UWash-00¹⁵, BIPM-01⁹, UWup-02⁴², MSL-03⁴³, HUST-05^{16,17}, UZur-06⁴⁴,

HUST-09^{18,19}, JILA-10⁴⁵, BIPM-14^{10,11}, LENS-14⁴⁷, UCI-14⁴⁶) and the CODATA-2014 value⁴. All error bars denote 1σ confidence level.

layer, and the correction is reduced to less than 2 p.p.m. owing to the low density of Al.

Air density

In the AAF method, the source masses are located in air, outside the vacuum chamber. The volume of air displaced by the sphere introduces a negative gravitational torque at the signal frequency. The associated correction to G is $\rho_{\text{air}}/\rho_{\text{sphere}}$, where $\rho_{\text{air}} \approx 1.18 \text{ kg m}^{-3}$ is the average air density, which is monitored by an air density measurement system, and $\rho_{\text{sphere}} \approx 7,965 \text{ kg m}^{-3}$ is the average sphere density. The average correction is 148.50 p.p.m. with an uncertainty of less than 1.51 p.p.m. In each run, the correction for this effect is applied in real time according to the measured air density. In the TOS method, both the pendulum and source masses are placed in the same vacuum chamber, thus no air density effect needs to be considered.

The thermal effect

In both methods, corrections were applied for thermal effects on all the geometrical parameters, such as the pendulum's dimensions and the distance between the geometric centres of the spheres. The torsion spring constant of the fibre is also temperature-dependent owing to thermoelasticity³⁴. For a small range of temperature variation, the spring constant of the fibre is linearly proportional to the temperature. The typical thermoelastic coefficient of the silica fibre used in this work was determined to be $101(1) \times 10^{-6} \text{ }^{\circ}\text{C}^{-1}$ using a temperature modulation experiment^{23,35,36}. This coefficient is slightly different from fibre to fibre. According to the monitored temperature variation around the fibre, the correction for the thermoelastic effect was applied synchronously for each run when extracting the oscillation frequency of the pendulum in the TOS method (Extended Data Table 5).

In the AAF method, the thermoelastic effect is negligible because the fibre does not twist. In addition, the temperature variation in the room was increased to about 1°C , and the response coefficient of angular acceleration of the pendulum turntable was measured to be $(2.2 \pm 3.6) \times 10^{-12} \text{ rad s}^{-2} \text{ }^{\circ}\text{C}^{-1}$ (Extended Data Fig. 4). Considering that the temperature variation was less than 0.1°C during each experimental run, it contributes an uncertainty of no more than 0.91 p.p.m.

The electrostatic effect

In the TOS method, the electrostatic disturbance was effectively reduced by the shield inserted between the pendulum and the source masses. During data acquisition, the pendulum, the shield and the source masses were all grounded. However, the fluctuation of the electrostatic potential difference between the shield and the pendulum could change the effective spring constant of the fibre and

affect the oscillation period. We measured the oscillation period of the pendulum for a varying voltage applied on the shield. The typical response coefficient of the period to the voltage was $-28.6(1) \text{ ms V}^{-1}$ near 0 V, corresponding to an extra electrostatic spring constant of $1.34(1) \times 10^{-12} \text{ N m rad}^{-1}$ per volt. When the spheres were exchanged between the 'near' and 'far' positions, the potential variation on the shield was measured by a digital multimeter to be less than $10 \mu\text{V}$, which contributes an uncertainty of no more than 0.17 p.p.m. to the G value. We applied different voltages on the shield in the sequence ground, 0.1 V, -0.1 V, ground, and found that the period of the pendulum changed correspondingly, but the period differences between the 'near' and 'far' positions were consistent with each other (Extended Data Fig. 5). This further confirms that the electrostatic effect on the G measurement with the TOS method is very small.

In the AAF method, a grounded vacuum chamber made of aluminium alloy shields electrostatically the grounded pendulum from the source masses. We found no substantial influence of the pendulum oscillation on the noise spectrum when a 1-mHz square wave voltage with an amplitude of about 10 V was applied on the upper-layer spheres (Fig. 1b).

The magnetic effect

In the TOS method, the interaction between the local magnetic field and residual magnetic moment of the spheres produces an additional torque on the pendulum. The contribution of this effect to the uncertainty of G was evaluated to be 2.08 p.p.m. (in TOS-I) and 0.71 p.p.m. (in TOS-II), following the method used in ref. ³⁷. In the AAF method, the horizontal magnetic gradient generated by the source masses produces a periodic torque on the pendulum at a signal frequency of $2\omega_d$. We measured this correction to be $24.2(1.4) \text{ p.p.m.}$ when an increased gradient of $0.31(1) \text{ G s m}^{-1}$ is produced by a current coil placed on the source-mass position. Because the background gradient induced by the four spheres is about 0.05 G s m^{-1} , the contribution to the uncertainty of G is less than 3.98 p.p.m. in AAF-I and AAF-II. In AAF-III, three layers of Mu-metal shields were used to enclose the pendulum, and this error was reduced to less than 0.90 p.p.m.

Data acquisition and analysis

In the TOS method, all the data, including the pendulum twist, the temperature, seismic disturbances and fluctuations of the air pressure, were taken at a regular intervals of 0.5 s triggered by a rubidium clock with a stability of 1×10^{-11} (at 1 s) and a frequency accuracy $\leq 1 \times 10^{-10}$. The data taking procedure for all experimental runs was the same as that used in our previous experiments^{18,19}. The acquisition time was three days for one position and the initial amplitude of the pendulum

oscillation was 3–4 mrad with an accuracy better than 56 μ rad. Typically, 10 sets of data were taken with the source masses in the two configurations alternately. The periods of the pendulum oscillation at the two configurations were extracted from the time-series angle data by the correlation method³⁸, and a typical result is shown in Fig. 2a. The thermoelastic and nonlinear properties of the fibre and the gravitational nonlinearity of the source masses were corrected synchronously (Supplementary Information Section 3 and Supplementary Table 4). In addition, the effects of the co-moving background gravitational gradient from the turntable and the supports were measured without the source masses following the above procedure. For each fibre, 10 sets of background data were collected (Fig. 2b), which were subtracted from the result obtained with the source masses in position.

In the AAF method, all data were taken at regular intervals of 1 s triggered by the same kind of rubidium clock as that used in the TOS method. In each interval, the data obtained in the first half second (Δt) were averaged and then saved in a computer during the second half of the interval. The pendulum turntable angle was numerically differentiated twice with a time increment of $\Delta T = 10$ s to yield the angular acceleration, a typical segment of which is shown in Fig. 2e. The true amplitude of the angular acceleration is attenuated by a factor of $\sin(\omega_d \Delta t)/(\omega_d \Delta t)$ and $[\sin(\omega_d \Delta T)/(\omega_d \Delta T)]^2$ owing to the data average in the first half second and the numeric derivative, respectively (Supplementary Information Section 4). The asymmetric mass distribution and imperfection of the ULE-glass shelf and the rotating parts of the source-mass turntable can generate a gravitational signal on the pendulum at the frequency of interest. To eliminate this co-moving background gravity gradient effect, we placed specially fabricated mass blocks on the shelf to compensate for the gravity gradient, and this effect was reduced to less than 2 p.p.m. (Supplementary Information Section 5 and Supplementary Fig. 1).

In AAF-I, four data sets were recorded, each of them 3–6 days long. In each run, the orientation of each sphere was changed by a random azimuthal angle to average out the density inhomogeneity effect of the source masses. The least-squares method was used to fit the angular acceleration data of the pendulum turntable, including the signal and its harmonics, the laboratory-fixed background and its harmonics, the linear drift and the offset. In AAF-II and AAF-III, the signal frequency $2\omega_d$ was changed from ~ 2.50 mHz (used in AAF-I) to ~ 1.67 mHz. In AAF-II and AAF-III, 10 and 15 sets of data were taken with different orientations of the spheres in each run, respectively. The G values determined from the three individual experiments are consistent, as shown in Fig. 2f.

Results

The systematic and statistical uncertainties are presented in Table 1. In the TOS method, fibres 1–3 and fibre 4 were used in TOS-I and TOS-II, respectively. Because the change in the period between two positions using fibre 2 is only 10% of that obtained when using other fibres, owing to the thicker diameter of fibre 2, a larger relative uncertainty of $\Delta\omega^2$ (the change of the squared frequency of the torsion pendulum with the source masses at the two configurations) is introduced. From 2014 to 2017, the G measurement was carried out once with fibre 2 and twice with random orientations of the source masses for fibres 1, 3 and 4. We obtained seven values of G for the four fibres (Fig. 2c and Supplementary Table 2). The weighted mean values of G for fibres 1, 2, 3 and 4 are $6.674187(91)G_0$, $6.674237(219)G_0$, $6.674269(93)G_0$ and $6.674061(104)G_0$ with relative uncertainties of 13.67, 32.88, 13.96 and 15.59 p.p.m., respectively, where $G_0 = 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$. These four results show good consistency within the relevant uncertainties. The correlations of the uncertainty components of the four results are discussed in Supplementary Information Section 6. Taking into account the correlation between the four fibres, the weighted mean value of G for the TOS method is $6.674184(78) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ with a combined relative uncertainty of 11.64 p.p.m. (1σ). The relative weights of the four G values are estimated as the reciprocal of the square of their uncertainties and are 0.345, 0.060, 0.330 and 0.265, respectively.

In the AAF method, the three experiments, AAF-I, AAF-II and AAF-III, give G values of $6.674534(83)G_0$, $6.674375(82)G_0$ and $6.674535(75)G_0$ with relative uncertainties of 12.45, 12.27 and 11.21 p.p.m., respectively (Supplementary Table 3). According to the method discussed above, the relative weights of these G values are estimated to be 0.306, 0.315 and 0.378, respectively. Taking into account the correlation between the three individual experiments (Supplementary Information Section 6), the weighted mean value of G for the AAF method is $6.674484(78) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ with a combined relative uncertainty of 11.61 p.p.m. (1σ).

Figure 3 shows a comparison of our results with the values of recent experiments^{9–11,15–19,39–47} and the CODATA-2014 adjustment⁴. It should be emphasized that different members of our group carried out the TOS-method and AAF-method experiments on different apparatus, so there is no correlation between the systematic errors of the two methods, to the best of our knowledge. The G values obtained with the two independent methods have the smallest uncertainty reported until now and both agree with the CODATA-2014 value within a 2σ range, indicating the substantial contribution of this work to the determination of the true value of G .

Furthermore, the value obtained here with the TOS method is larger than our previous measurement (HUST-09^{18,19} in Fig. 3) by more than a hundred parts per million, but we currently have no definite explanation for the inconsistency between the two results (Supplementary Information Section 7). This illustrates that determining the true value of G is very difficult, and further measurements are needed in the future.

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0431-5>.

Received: 29 April 2018; Accepted: 5 July 2018;

Published online 29 August 2018.

- Cavendish, H. Experiments to determine the density of the Earth. *Phil. Trans. R. Soc. B* **88**, 469–526 (1798).
- Gillies, G. T. The Newtonian gravitational constant: an index of measurements. *Metrologia* **24**, 1–56 (1987).
- Rothleitner, C. & Schlamminger, S. Measurements of the Newtonian constant of gravitation. *G. Rev. Sci. Instrum.* **88**, 111101 (2017).
- Mohr, P. J., Newell, D. B. & Taylor, B. N. CODATA recommended values of the fundamental physical constants: 2014. *Rev. Mod. Phys.* **88**, 035009 (2016).
- Quinn, T. Measuring big G . *Nature* **408**, 919–921 (2000).
- Quinn, T. Don't stop the quest to measure Big G . *Nature* **505**, 455 (2014).
- Schlamminger, S. Fundamental constants: a cool way to measure big G . *Nature* **510**, 478–480 (2014).
- Gibney, E. Rivals join forces to nail down Big G . *Nature* **514**, 150–151 (2014).
- Quinn, T. J., Speake, C. C., Richman, S. J., Davis, R. S. & Picard, A. A new determination of G using two methods. *Phys. Rev. Lett.* **87**, 111101 (2001).
- Quinn, T. J., Parks, H. V., Speake, C. C. & Davis, R. S. Improved determination of G using two methods. *Phys. Rev. Lett.* **111**, 101102 (2013).
- Quinn, T., Speake, C., Parks, H. & Davis, R. The BIPM measurements of the Newtonian constant of gravitation. *G. Phil. Trans. R. Soc. A* **372**, 20140032 (2014).
- Heyl, P. R. A redetermination of the constant of gravitation. *J. Res. Natl. Bur. Stand.* **5**, 1243–1290 (1930).
- Heyl, P. R. & Chrzanowski, P. A new determination of the constant of gravitation. *J. Res. Natl. Bur. Stand.* **29**, 1–31 (1942).
- Rose, R. D., Parker, H. M., Lowry, R. A., Kuhlthau, A. R. & Beams, J. W. Determination of the gravitational constant G . *Phys. Rev. Lett.* **23**, 655–658 (1969).
- Gundlach, J. H. & Merkowitz, S. M. Measurement of Newton's constant using a torsion balance with angular acceleration feedback. *Phys. Rev. Lett.* **85**, 2869–2872 (2000).
- Luo, J. H., Hu, Z. K., Fu, X. H., Fan, S. H. & Tang, M. X. Determination of the Newtonian gravitational constant G with a nonlinear fitting method. *Phys. Rev. D* **59**, 042001 (1998).
- Hu, Z. K., Guo, J. Q. & Luo, J. Correction of source mass effects in the HUST-99 measurement of G . *Phys. Rev. D* **71**, 127505 (2005).

18. Luo, J. et al. Determination of the Newtonian gravitational constant G with time-of-swing method. *Phys. Rev. Lett.* **102**, 240801 (2009).
19. Tu, L. C. et al. New determination of the gravitational constant G with time-of-swing method. *Phys. Rev. D* **82**, 022001 (2010).
20. Li, Q. et al. G measurements with time-of-swing method at HUST. *Phil. Trans. R. Soc. A* **372**, 20140141 (2014).
21. Kuroda, K. Does the time-of-swing method give a correct value of the Newtonian gravitational constant? *Phys. Rev. Lett.* **75**, 2796–2798 (1995).
22. Newman, R. D. & Bantel, M. K. On determining G using a cryogenic torsion pendulum. *Meas. Sci. Technol.* **10**, 445–453 (1999).
23. Yang, S. Q. et al. Direct measurement of the anelasticity of a tungsten fiber. *Phys. Rev. D* **80**, 122005 (2009).
24. Xue, C. et al. Preliminary determination of Newtonian gravitational constant with angular acceleration feedback method. *Phil. Trans. R. Soc. A* **372**, 20140031 (2014).
25. Quan, L. D. et al. Feedback control of torsion balance in measurement of gravitational constant G with angular acceleration method. *Rev. Sci. Instrum.* **85**, 014501 (2014).
26. Fan, X. D. et al. Coupled modes of the torsion pendulum. *Phys. Lett. A* **372**, 547–552 (2008).
27. Numata, K., Horowitz, J. & Camp, J. Coated fused silica fibers for enhanced sensitivity torsion pendulum for LISA. *Phys. Lett. A* **370**, 91–98 (2007).
28. Li, Q. et al. Research on supporting mounts of spheres in measurement of gravitational constant G . *Rev. Sci. Instrum.* **87**, 034504 (2016).
29. Luo, J., Wang, W. M., Hu, Z. K. & Wang, X. L. Precise determination of separation between spherical attracting masses in measuring the gravitational constant. *Chin. Phys. Lett.* **18**, 1012–1014 (2001).
30. Liu, L. X. et al. Measurement of density inhomogeneity for glass pendulum. *Chin. Phys. Lett.* **25**, 4203–4206 (2008).
31. Liu, L. X., Shao, C. G., Tu, L. C. & Luo, J. Measurement of density inhomogeneity for source masses in time-of-swing method of measuring G . *Chin. Phys. Lett.* **26**, 010403 (2009).
32. Guo, J. Q., Hu, Z. K., Gu, B. M. & Luo, J. Measurement of eccentricity of the centre of mass from the geometric centre of a sphere. *Chin. Phys. Lett.* **21**, 612–615 (2004).
33. Liu, L. X. et al. Precision measurement of distribution of film thickness on pendulum for experiment of G . *Chin. Phys. Lett.* **26**, 090402 (2009).
34. Zener, C. *Elasticity and Anelasticity of Metals* (University of Chicago Press, Chicago, 1948).
35. Luo, J., Hu, Z. K. & Hsu, H. Thermoelastic property of the torsion fiber in the gravitational experiments. *Rev. Sci. Instrum.* **71**, 1524–1528 (2000).
36. Hu, Z. K., Wang, X. L. & Luo, J. Thermoelastic correction in the torsion pendulum experiment. *Chin. Phys. Lett.* **18**, 7–9 (2001).
37. Li, Q., Liu, L. X., Tu, L. C., Shao, C. G. & Luo, J. Effect of local magnetic field in G measurement with time-of-swing method. *Chin. Phys. Lett.* **27**, 070401 (2010).
38. Tian, Y. L., Tu, Y. & Shao, C. G. Correlation method in period measurement of a torsion pendulum. *Rev. Sci. Instrum.* **75**, 1971–1974 (2004).
39. Luther, G. G. & Towler, W. R. Redetermination of the Newtonian gravitational constant G . *Phys. Rev. Lett.* **48**, 121–123 (1982).
40. Karagioz, O. & Izmailov, V. Measurement of the gravitational constant with a torsion balance. *Meas. Tech.* **39**, 979–987 (1996).
41. Bagley, C. H. & Luther, G. G. Preliminary results of a determination of the Newtonian constant of gravitation: a test of the Kuroda hypothesis. *Phys. Rev. Lett.* **78**, 3047–3050 (1997).
42. Kleinevoss, U. Bestimmung der Newtonschen Gravitationskonstanten G . PhD thesis (Univ. Wuppertal, 2002); <http://elpub.bib.uni-wuppertal.de/servlets/DocumentServlet?id=335&lang=en>.
43. Armstrong, T. R. & Fitzgerald, M. P. New measurements of G using the measurement standards laboratory torsion balance. *Phys. Rev. Lett.* **91**, 201101 (2003).
44. Schlamminger, S. et al. Measurement of Newton's gravitational constant. *Phys. Rev. D* **74**, 082001 (2006).
45. Parks, H. V. & Faller, J. E. Simple pendulum determination of the gravitational constant. *Phys. Rev. Lett.* **105**, 110801 (2010).
46. Newman, R., Bantel, M., Berg, E. & Cross, W. A measurement of G with a cryogenic torsion pendulum. *Phil. Trans. R. Soc. A* **372**, 20140025 (2014).
47. Rosi, G., Sorrentino, F., Cacciapuoti, L., Prevedelli, M. & Tino, G. M. Precision measurement of the Newtonian gravitational constant using cold atoms. *Nature* **510**, 518–521 (2014).

Acknowledgements We are grateful to R. Newman, T. Quinn, C. Speake, J. E. Faller, J. H. Gundlach, H. J. Paik, Z. H. Lu, J. Luo and S. H. Fan for discussions and suggestions. We thank Q. T. Fan, Y. T. Zhang, B. P. Wang, X. D. Fan, M. Ke, L. Zhao, Y. Tu, J. Q. Guo, D. C. Chen, W. M. Wang, X. L. Wang, X. J. Luo, X. H. Fu, J. Tang and Y. B. Cheng for their early works on G measurement. We thank the National Institute of Metrology (NIM) of China for the calibration of some measuring instruments, source masses and the length gauges. This work is partly supported by the National Natural Science Foundation of China under grants number 91536223, 11722542, 11325523 and 11605295, the National Basic Research Program of China under grant number 2010CB832801 and the National Precise Gravity Measurement Facility.

Reviewer information *Nature* thanks S. Schlamminger and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.L. had the idea for the experiment. J.L. and S.-Q.Y. supervised all the experiments. Q.Li and J.-P.L. performed the experiment with the TOS method and analysed the data. C.X. and J.-F.W. performed the experiment with the AAF method and analysed the data. L.-D.Q. designed and built the feedback control system of the two turntables in the AAF method. C.-G.S. analysed all the errors and data independently. W.-H.T., H.X., L.-C.T., Q.Liu, L.-X.L., Q.-L.W., Z.-K.H., Z.-B.Z., P.-S.L., S.-C.W. and V.M. contributed to the analysis and discussion. S.-Q.Y., Q.Li and C.X. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

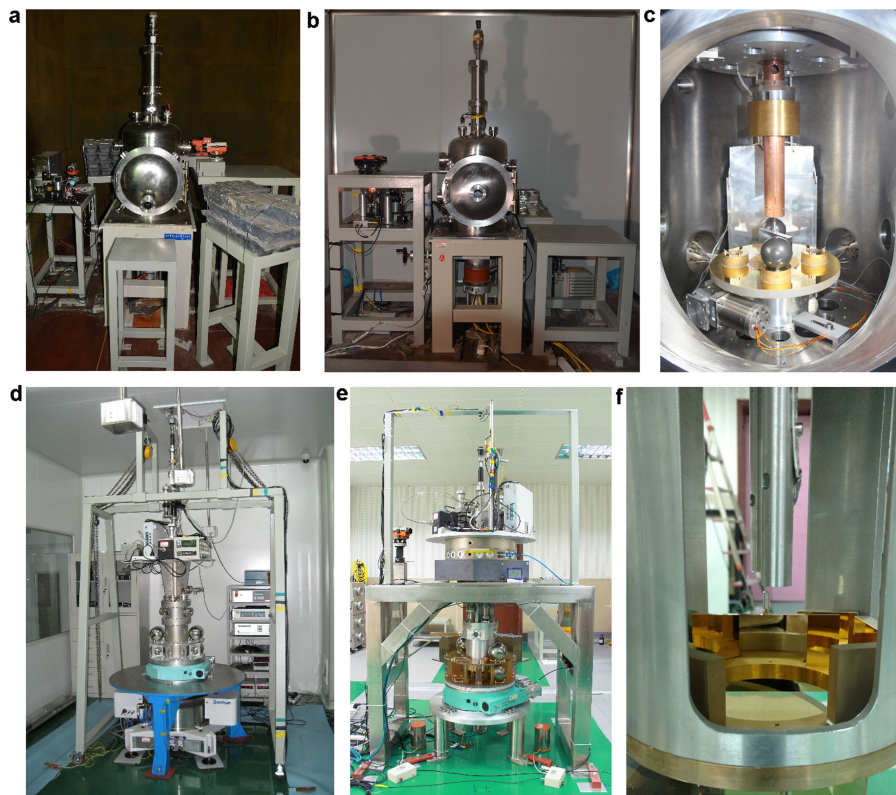
Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0431-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0431-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

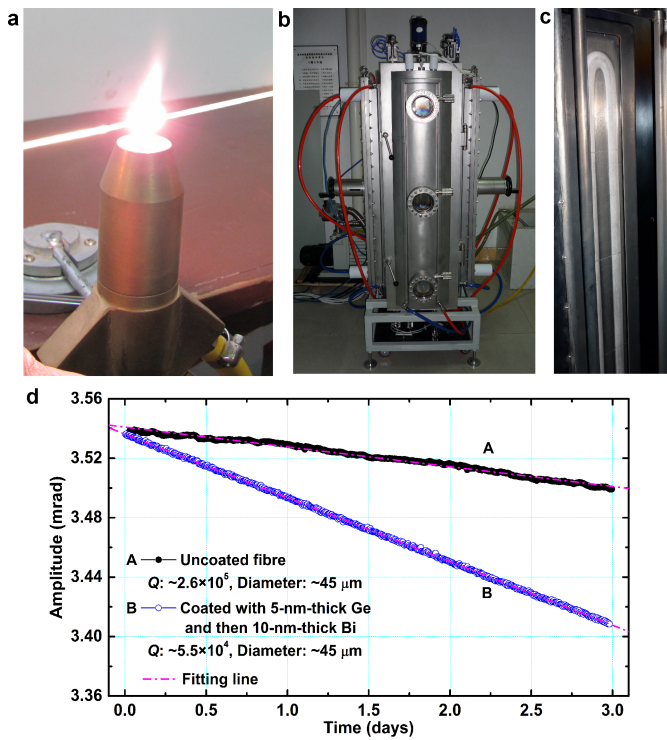
Correspondence and requests for materials should be addressed to S.-Q.Y., C.-G.S. and J.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

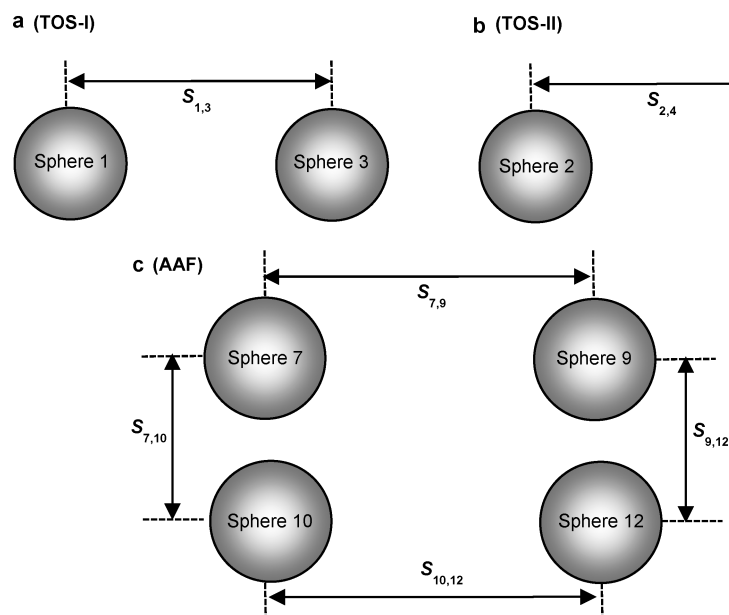


Extended Data Fig. 1 | Photographs of the experimental apparatus. **a**, Apparatus 1, used in TOS-I. **b**, Apparatus 2, used in TOS-II. **c**, The suspended pendulum and source masses in the vacuum chamber used in the TOS method. The copper tube around the fibre is used to reduce the temperature gradient. The electrostatic shield (here elevated to show the pendulum), the three-point mounts, the ULE-glass disk and the

turntable are also shown. **d**, The preliminary apparatus used to perform the proof-of-principle measurements^{24,25} of G using the AAF method. **e**, The improved apparatus used in the present work. The apparatus was completely rebuilt to reduce several sources of uncertainty encountered in the proof-of-principle experiments (see text for details). **f**, The suspended pendulum and the optical path system used in the AAF method.

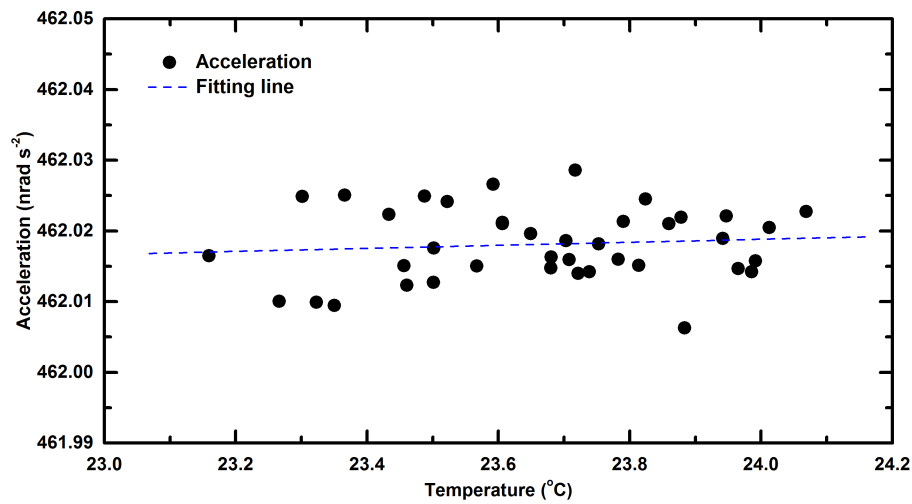


Extended Data Fig. 2 | Fabrication of the silica fibre and measurement of its Q factor. **a**, Photograph of a silica fibre pulled from a rod over an oxygen-natural gas flame. **b**, Magnetron sputtering equipment used for the coating of the silica fibres. **c**, The Bi target, with a height of ~1 m. The Ge target (not shown here) is similar. The two targets are installed on opposite sides of the coating equipment, with the fibre located between the two targets and rotated continuously. The surfaces of the fibres were coated with a 5-nm-thick Ge layer and then a 10-nm-thick Bi layer. **d**, Typical decay curves of the torsional amplitude of a pendulum suspended by a ~45-μm-diameter fibre. Curve A represents the uncoated silica fibre, with a Q factor of 2.6×10^5 . Curve B corresponds to the coated silica fibre, with a Q factor of 5.5×10^4 . The dot-dashed lines denote fitting curves of the exponential function $A = A_0 \exp(-\pi f_0 t/Q)$, where A_0 is the initial amplitude, f_0 is the free oscillation frequency and t is the time.



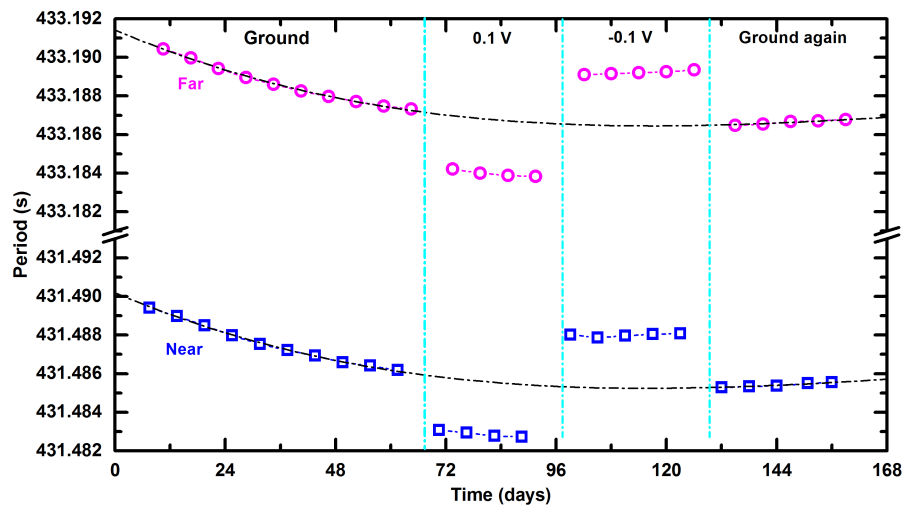
Extended Data Fig. 3 | Schematic diagram of the source masses. a, b, In the TOS method, spheres 1 and 3 are used in apparatus 1 (TOS-I; **a**) and spheres 2 and 4 are used in apparatus 2 (TOS-II; **b**). $S_{1,3}$ and $S_{2,4}$ are the horizontal distances of the geometric centres of the spheres in apparatus

1 and 2, respectively. **c,** In the AAF method, spheres 7, 9, 10 and 12 are used. $S_{7,9}$ and $S_{10,12}$ are the horizontal distances and $S_{7,10}$ and $S_{9,12}$ are the vertical distances between the geometric centres of the spheres.



Extended Data Fig. 4 | Effect of temperature on the measurement of the angular acceleration in the AAF method. A modulation experiment was carried out by increasing the temperature variation in the room to about 1 °C. Solid circles represent the average angular acceleration of the pendulum turntable over 12-h data taking periods. The dashed line with

a slope of $(2.2 \pm 3.6) \times 10^{-12} \text{ rad s}^{-2} \text{ }^{\circ}\text{C}^{-1}$ represents the least-squares fitting curve. The result indicates that the apparatus is insensitive to the temperature variation and that a temperature variation of less than 0.1 °C during each experimental run contributes an uncertainty of less than 0.91 p.p.m. to the G measurement.



Extended Data Fig. 5 | Electrostatic effect on the measurement of the pendulum period in the TOS method. Different voltages are applied on the shield in the sequence: ground, 0.1 V, −0.1 V, ground. For each voltage, 4–5 sets of measurements of the pendulum period are performed at the ‘near’ and ‘far’ configurations. The corresponding change of the frequency squared ($\Delta\omega^2$) for the steps of the sequence is determined to be

$1.662192(8) \times 10^{-6} \text{ s}^{-2}$, $1.662184(16) \times 10^{-6} \text{ s}^{-2}$, $1.662181(15) \times 10^{-6} \text{ s}^{-2}$ and $1.662200(13) \times 10^{-6} \text{ s}^{-2}$, respectively. The results show that the period changes with the applied voltage, but the $\Delta\omega^2$ values for the ‘near’ and ‘far’ configurations are consistent with each other within the statistical uncertainty. The dot-dashed lines are polynomial fitting curves that represent the period drift due to the ‘aging’ effect of the fibre.

Extended Data Table 1 | Dimensions and masses of the pendulums

Parameters		Length (mm)	Width (mm)	Height (mm)	Mass (g)
TOS method	Pendulum 1	91.00575(11)	11.08688(9)	30.66846(12)	68.09937(22)
	Pendulum 2	91.00336(17)	11.04448(13)	30.00446(13)	66.35715(22)
AAF method	Pendulum 3	91.05243(29)	4.00240(8)	49.92441(24)	40.0379(3)

In the TOS method, pendulum 1 is used in apparatus 1, and all the dimensions are converted to the values at 20.2 °C. Pendulum 2 is utilized in apparatus 2, and all the dimensions are converted to the values at 21.5 °C. Pendulum 3 is used in the AAF method, and all the dimensions are converted to the values at 23.7 °C. The temperature is the average value over the data acquisition period in each measurement of G. Uncertainties are one standard deviation.

Extended Data Table 2 | Parameters of the source masses

Parameters		Diameter (mm)	Mass (g)	Nonsphericity (μm)
TOS-I	Sphere 1	57.15072(25)	778.1630(8)	0.22(3)
	Sphere 3	57.14577(25)	777.9649(8)	0.23(3)
TOS-II	Sphere 2	57.15236(30)	778.1789(6)	0.23(3)
	Sphere 4	57.15187(31)	778.1754(6)	0.27(9)
AAF method (AAF-I, II, III)	Sphere 7	127.0003(8)	8,543.5826(53)	0.75(6)
	Sphere 9	126.9957(9)	8,541.4167(53)	0.75(6)
	Sphere 10	126.9934(8)	8,540.5282(52)	0.72(6)
	Sphere 12	126.9887(10)	8,541.5575(53)	0.89(11)

In the TOS method, spheres 1 and 3 are used in TOS-I, and all the dimensions are converted to the values at 20.2 °C. Spheres 2 and 4 are utilized in TOS-II, and all the dimensions are converted to the values at 21.5 °C. Spheres 7, 9, 10 and 12 are used in the AAF method, and all the dimensions are converted to the values at 23.7 °C. Uncertainties are one standard deviation.

Extended Data Table 3 | Comparison of several main corrections between the current experiment and our previous experiment^{18,19}

Item	TOS method (in units of p.p.m.)					AAF method (in units of p.p.m.)		
	HUST-09	TOS-I: Fibre 1	TOS-I: Fibre 2	TOS-I: Fibre 3	TOS-II: Fibre 4	AAF-I	AAF-II	AAF-III
Coating layer	-24.28 (4.33)	-1.70(86)	-1.70(86)	-1.70(86)	-1.52(73)	-9.10(34)	-9.09(34)	-9.10(34)
Clamp	1,297.29 (1.62)	70.66(14)	70.65(14)	70.73(14)	68.58(33)	5.75(15)	5.85(21)	5.73(11)
Ferrule	105.22(30)	12.10(5)	12.91(5)	12.70(6)	12.85(5)	22.52(69)	22.84 (1.03)	22.45(47)
Others	11.98(21)	9.27(40)	10.19(37)	9.51(39)	8.79(26)	3.23(29)	3.29(29)	3.21(29)
Fibre anelasticity	-211.80 (18.69)	-6.01 (3.00)	-8.38 (4.19)	-5.68 (2.84)	-6.92 (3.46)	0.01	0.01	0.01
Magnetic damper	17.54(31)	0.47(8)	7.13 (1.19)	0.32(5)	0.27(8)	455.40 (1.95)	455.40 (1.95)	25.74(8)
Average Air density effect	—	—	—	—	—	149.90 (1.00)	147.33 (1.51)	148.27 (1.13)
Data averaging $\Delta t=0.5$ s	—	—	—	—	—	2.57(1)	1.14(1)	1.14(1)
Numeric derivatives $\Delta T=10$ s	—	—	—	—	—	2,058.71 (1)	914.35(1)	914.35(1)

Coating layer: in the current experiment with the TOS method, the effect of the coating layer is reduced by choosing aluminium as the coating material to replace Au/Cu, which was used in a previous experiment (HUST-09)^{18,19}. Clamp and ferrule: in the current experiment with the TOS method, the aluminium clamp and ferrule that used to connect the pendulum and the silica fibre in the previous experiment are miniaturized. The corresponding corrections are reduced to 1/18 and 1/8 of those in HUST-09, respectively. 'Others' includes effects due to the pendulum mass, the reflecting mirror, glues, edge flaws and the silica rod in both methods. Fibre anelasticity: this effect is reduced by choosing the high- Q silica fibre to replace the tungsten fibre used in HUST-09. Magnetic damper: this effect is reduced when the prehanger fibre is shorter and thicker. Data averaging and numerical derivatives: the true amplitude of the angular acceleration of the pendulum turntable is attenuated by a factor of $[\sin(\omega_d \Delta t)]/(\omega_d \Delta t)$ and $\{[\sin(\omega_d \Delta T)]/(\omega_d \Delta T)\}^2$ owing to averaging in the data acquisition and the use of numerical derivatives in data processing, respectively (see Supplementary Information Section 4). Values in parentheses are the uncertainties of the corrections. Uncertainties are one standard deviation.

Extended Data Table 4 | Distance between the geometric centres of the spheres

Items		Temperature (°C)	GC Distance (mm)
TOS-I: Fibre 1	First experiment	20.3	157.19245(34)
	Repeated experiment	20.3	157.19363(34)
TOS-I: Fibre 2		20.3	157.19363(34)
TOS-I: Fibre 3	First experiment	20.1	157.19392(33)
	Repeated experiment	20.1	157.19384(33)
TOS-II: Fibre 4	First experiment	21.5	157.16476(37)
	Repeated experiment	21.5	157.16489(36)
AAF method (AAF-I, II, III)	Between Sphere 7 and 9 ($S_{7,9}$)	23.7	342.2874(19)
	Between Sphere 10 and 12 ($S_{10,12}$)	23.7	342.3074(19)
	Between Sphere 7 and 10 ($S_{7,10}$)	23.7	139.7997(15)
	Between Sphere 9 and 12 ($S_{9,12}$)	23.7	139.7822(17)

In the TOS method, the temperature coefficient is measured to be less than $0.11 \mu\text{m } ^\circ\text{C}^{-1}$. The temperature variation is less than $0.1 ^\circ\text{C}$ during each experimental run, which contributes an uncertainty of 0.30 p.p.m. In the AAF method, the temperature coefficient of the horizontal geometric centre (GC) distance of the upper-layer spheres is $-1.9(1) \mu\text{m } ^\circ\text{C}^{-1}$, which is used to correct the geometric centre distances. The lower horizontal and the vertical geometric centre distances are found to be constant within an uncertainty of $2 \mu\text{m}$ for a temperature change of $4 ^\circ\text{C}$. Uncertainties are one standard deviation.

Extended Data Table 5 | Thermoelastic effect corrections for each fibre used in the TOS method

Items		Thermoelastic effect (in units of p.p.m.)	Average temperature (°C)
TOS-I: Fibre 1	First experiment	-73.13(0.71)	20.3
	Repeated experiment	65.08(0.71)	20.3
TOS-I: Fibre 2		723.89(3.41)	20.3
TOS-I: Fibre 3	First experiment	-146.56(0.77)	20.1
	Repeated experiment	-85.51(0.61)	20.1
TOS-II: Fibre 4	First experiment	-90.92(0.97)	21.5
	Repeated experiment	-145.59(1.46)	21.5

Uncertainties are one standard deviation.

5-HT release in nucleus accumbens rescues social deficits in mouse autism model

Jessica J. Walsh¹, Daniel J. Christoffel¹, Boris D. Heifets², Gabriel A. Ben-Dor¹, Aslihan Selimbeyoglu^{1,3,4}, Lin W. Hung¹, Karl Deisseroth^{3,4} & Robert C. Malenka^{1*}

Dysfunction in prosocial interactions is a core symptom of autism spectrum disorder. However, the neural mechanisms that underlie sociability are poorly understood, limiting the rational development of therapies to treat social deficits. Here we show in mice that bidirectional modulation of the release of serotonin (5-HT) from dorsal raphe neurons in the nucleus accumbens bidirectionally modifies sociability. In a mouse model of a common genetic cause of autism spectrum disorder—a copy number variation on chromosome 16p11.2—genetic deletion of the syntenic region from 5-HT neurons induces deficits in social behaviour and decreases dorsal raphe 5-HT neuronal activity. These sociability deficits can be rescued by optogenetic activation of dorsal raphe 5-HT neurons, an effect requiring and mimicked by activation of 5-HT_{1b} receptors in the nucleus accumbens. These results demonstrate an unexpected role for 5-HT action in the nucleus accumbens in social behaviours, and suggest that targeting this mechanism may prove therapeutically beneficial.

Positive prosocial interactions contribute to the development and maintenance of a range of adaptive, cooperative behaviours. Conversely, abnormal social interactions are debilitating symptoms of several neuropsychiatric disorders, notably autism spectrum disorder (ASD)^{1,2}. Although the role of neuromodulators in social behaviours, in particular oxytocin, is an active area of investigation, relatively little is known about the neural mechanisms that influence sociability. Activation of oxytocin receptors on 5-HT terminals in the nucleus accumbens (NAc) is essential for the processing of social rewards³, raising the possibility that 5-HT release in the NAc contributes to prosocial interactions. The serotonergic system has long been implicated in behavioural deficits associated with psychiatric disorders^{4,5}, with some findings suggesting that changes in brain levels of 5-HT or manipulations of 5-HT signalling can influence social behaviours^{6–10}. Indeed, an early clue for a role of 5-HT in social behaviour came from measuring abnormal blood 5-HT levels in children with autism¹¹.

Here, we directly test the hypothesis that bidirectional modulation of 5-HT release in the NAc bidirectionally modifies sociability. We study the importance of this mechanism for ASD by examining a mouse model of copy number variations on human chromosome 16p11.2, a common genetic variation associated with ASD^{12–14}. Deletion of chromosome 7F3, which is syntenic to human 16p11.2, specifically from 5-HT neurons induced deficits in social behaviour, decreased dorsal raphe (DR) 5-HT activity during social contact, and reduced DR 5-HT neuron excitability. The decrease in sociability in 16p11.2 deletion mice was rescued by activation of DR 5-HT neurons, an effect requiring NAc 5-HT_{1b} receptors. Similar rescue of social behaviour deficits was achieved by pharmacological activation of NAc 5-HT_{1b} receptors. These results establish the importance of 5-HT action in the NAc in prosocial behaviours and suggest a rational path for the development of new therapeutic agents for the treatment of social behaviour deficits in neuropsychiatric disorders.

NAc 5-HT terminal activity influences sociability

We examined how DR neuron activity regulates social behaviour by injecting adeno-associated viruses (AAV) expressing channelrhodopsin-2

fused to enhanced yellow fluorescent protein (ChR2-eYFP) or eYFP alone into the DR of wild-type mice implanted with an optic fibre above the DR (Fig. 1a). Sociability was assayed by juvenile interaction and three-chamber tests along with open field locomotion and novel object interaction assays (Fig. 1b). Activation of DR neurons expressing ChR2 (at 20 Hz) caused increases in sociability in both assays, whereas no effects were observed in eYFP-expressing control mice that received light stimulation (Fig. 1c, d; Extended Data Fig. 1a, b). DR neuron activation had no effect on control behaviours (Fig. 1e, f) or anxiety-related behaviours (Extended Data Fig. 1c). Importantly, all assays were conducted and analysed in a blinded fashion.

The NAc contributes to the regulation of social behaviours^{3,15–19}. To determine whether the NAc was a crucial target of the activated DR neurons, we injected AAVs into the DR of wild-type mice and implanted optic fibres bilaterally above the NAc (Fig. 1g). Activation of DR terminals in the NAc expressing ChR2 caused increases in both sociability assays, whereas control mice exhibited no change in sociability (Fig. 1h–j; Extended Data Fig. 1d, e). Terminal activation of neurons projecting from the DR to the NAc did not influence the novel object interaction assay, locomotor activity or anxiety-related behaviours (Fig. 1k, l; Extended Data Fig. 1f).

5-HT is implicated in the regulation of social behaviours^{3,6–11} and a major source is DR 5-HT neurons²⁰. To test whether modulation of DR 5-HT neuron activity alters sociability, we injected double-floxed AAV-DIO-ChR2-eYFP (DIO-ChR2) or AAV-DIO-NpHR-eYFP (DIO-NpHR), which expresses the inhibitory opsin NpHR3.0, into the DR of mice that express Cre specifically in 5-HT neurons (*Sert-cre*, also known as *Slc6a4-cre*, mice)²¹ and implanted optic fibres above the DR (Fig. 2a, b). *Sert-cre* mice injected with AAV-DIO-eYFP served as controls. Optogenetic activation of DR 5-HT neurons increased sociability (Fig. 2c, d; Extended Data Fig. 1g, h), whereas optogenetic inhibition of DR 5-HT neurons decreased sociability (Fig. 2e, f; Extended Data Fig. 1i, j). Control mice were unaffected by light stimulation (Fig. 2c–f; Extended Data Fig. 1g–j) and control behaviours were unaffected by manipulation of DR 5-HT neuron activity (Extended Data Fig. 1k–p).

¹Nancy Pritzker Laboratory, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. ²Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA, USA. ³Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ⁴Department of Bioengineering and Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. *e-mail: malenka@stanford.edu

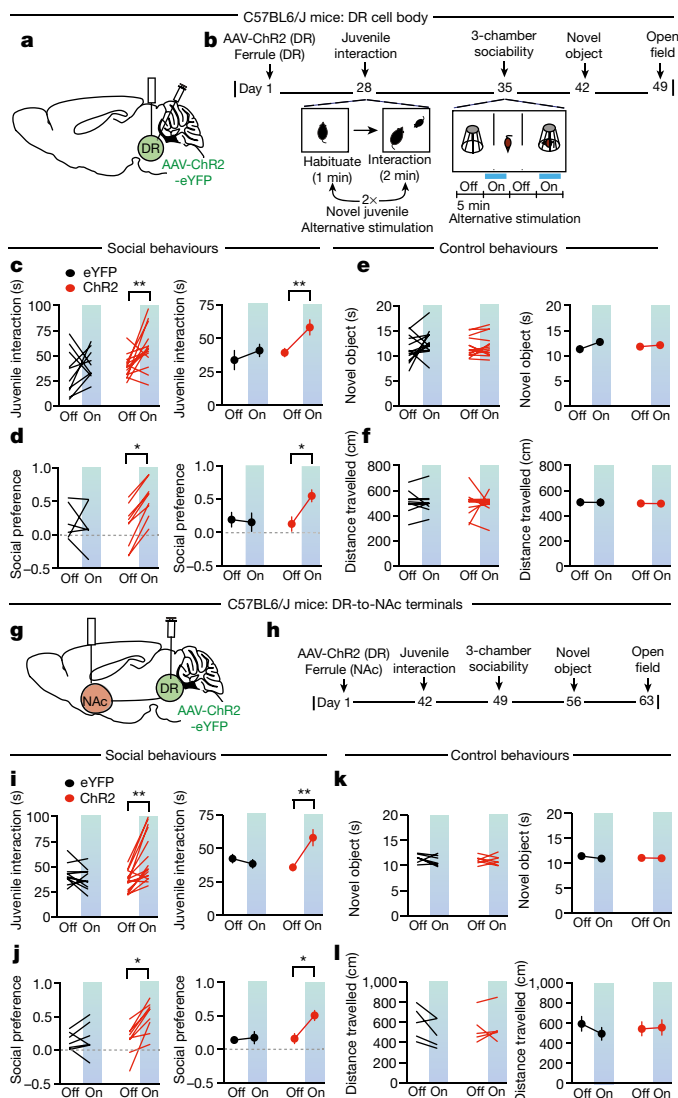


Fig. 1 | Activation of DR neurons or their NAc projections increases sociability. **a**, Schematic of optogenetic manipulation. **b**, Timeline of experiments. **c**, **d**, Quantification of juvenile interaction assay (**c**: $F_{1,44} = 6.262$, $P < 0.05$; $n = 10-14$) and three-chamber sociability assay (**d**: $F_{1,26} = 4.327$, $P < 0.05$; $n = 6-9$) in wild-type mice (blue signifies optical activation). In this and all subsequent figures, the left panels illustrate individual subjects; right panels display mean \pm s.e.m. **e**, **f**, DR neuron stimulation did not alter the novel object interaction assay (**e**: $F_{1,52} = 0.8324$, $P = 0.3658$, $n = 14$) or the locomotion assay (**f**: $F_{1,44} = 0.00053$, $P = 0.9817$; $n = 10-14$). **g**, Schematic of optogenetic manipulation. **h**, Timeline of experiments. **i**, **j**, Quantification of juvenile interaction (**i**: $F_{1,50} = 8.999$, $P < 0.01$; $n = 11-16$) and three-chamber sociability (**j**: $F_{1,28} = 4.320$, $P < 0.05$; $n = 7-9$) assays in wild-type mice. **k**, **l**, Stimulation of DR-to-NAc terminals does not alter the novel object interaction assay (**k**: $F_{1,24} = 0.4047$, $P = 0.5307$, $n = 7$) or the locomotion assay (**l**: $F_{1,16} = 0.6484$, $P = 0.4325$, $n = 5$). * $P < 0.05$, ** $P < 0.01$; two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

To test whether modifying 5-HT terminal activity in the NAc specifically recapitulates the consequences of DR 5-HT somatic manipulations, we expressed DIO-ChR2 or DIO-NpHR in the DR of *Sert-cre* mice and implanted optic fibres in the NAc (Fig. 2g, h). Identical to somatic stimulation, terminal activation of 5-HT neurons projecting from the DR to the NAc increased sociability (Fig. 2i, j; Extended Data Fig. 2a, b), whereas inhibition substantially decreased sociability (Fig. 2k, l; Extended Data Fig. 2c, d) with no effects in control mice

(Fig. 2i-l; Extended Data Fig. 2a-d) and no effects of ChR2 or NpHR activation on the novel object interaction assay, locomotor activity, or anxiety-related behaviours (Extended Data Fig. 2e-j).

Effects of 16p11.2 deletion in DR 5-HT neurons

Abnormalities in the brain's 5-HT system are implicated in ASD^{7,10,11,22}, providing motivation to test the relevance of our findings to ASD pathophysiology. We studied a mouse model of the 16p11.2 deletion syndrome because this is a common genetic variation associated with ASD¹²⁻¹⁴ and a floxed mouse line (*16p11.2^{flx}*) was available²³, allowing for control over deletion of the syntenic region on mouse chromosome 7F3 (Fig. 3a). We first examined whole brain 16p11.2 deletion by crossing *16p11.2^{flx}* mice to a *Nes-creER* mouse line (*16p11.2^{flx}:Nes-creER*), which limits Cre expression to neurons and is tamoxifen dependent allowing for temporal control of the genetic deletion²⁴ (Fig. 3b, c). Homozygous *16p11.2^{flx}:Nes-creER* mice exhibited decreased sociability, whereas heterozygous mice exhibited a decrease in juvenile interactions and a trend towards a decrease in three-chamber assays (Fig. 3d, e; Extended Data Fig. 3a, b). Neither group of 16p11.2 deletion mice exhibited changes in the novel object interaction assay (Fig. 3f), whereas homozygous *16p11.2^{flx}:Nes-creER* mice exhibited hyperactivity (Fig. 3g), as observed previously^{23,25}. None of the groups exhibited changes in anxiety-related behaviours in the open field test (Extended Data Fig. 3c).

To examine 16p11.2 deletion specifically in DR neurons as well as isolate the deletion solely to 5-HT neurons, we infused AAV-DJ-Cre²⁶ into the DR of *16p11.2^{flx}* mice (Fig. 3h) or crossed *Sert-cre* mice to floxed mice (*Sert-cre:16p11.2^{flx}*) (Fig. 3i, j). Both *16p11.2^{flx}* mice expressing Cre in the DR and *Sert-cre:16p11.2^{flx}* mice exhibited decreases in sociability compared to *16p11.2^{flx}* mice that had control AAV-DJ-ΔCre infused into DR (Fig. 3k, l; Extended Data Fig. 3d, e), with no abnormalities in the novel object interaction assay, locomotion, or anxiety-related behaviours (Fig. 3m, n; Extended Data Fig. 3f).

To assess DR 5-HT neuron activity during social interaction, *Sert-cre* and *Sert-cre:16p11.2^{flx}* mice were infused with AAV-DJ-DIO-GCaMP6f into the DR and fibre optics implanted above DR to perform fibre photometry recordings (Fig. 4a). DR 5-HT neuron activity increased during social interaction in *Sert-cre* mice⁹ and the magnitude of this increase was reduced in mice with 16p11.2 deleted from 5-HT neurons (Fig. 4b, c). To further assess how 16p11.2 deletion affects the function of DR 5-HT neurons, *Sert-cre* and *Sert-cre:16p11.2^{flx}* mice were injected with DIO-eYFP into the DR so that we could make whole-cell recordings from identified 5-HT neurons in acute DR slices. DR 5-HT neurons lacking 16p11.2 exhibited a decrease in spiking in response to depolarizing current pulses (Fig. 4d, e). These neurons also exhibited an approximately 50% decrease in the amplitude of spontaneous excitatory postsynaptic currents (Fig. 4f, g) with no change in their frequency (Fig. 4h, i). These results suggest that DR 5-HT neuron spiking is reduced after deletion of 16p11.2 and that impairment of DR 5-HT neuron activity contributes to the behavioural deficits observed in 16p11.2 deletion mice.

NAc 5-HT rescues social deficits in *16p11.2^{flx}* mice

If sociability deficits in 16p11.2 deletion mice are due to reduced function of DR 5-HT neurons, it may be possible to rescue the deficits by driving DR 5-HT neuron activity. To test this prediction, we infused AAV-DJ-Cre and either DIO-ChR2 or DIO-eYFP into the DR of *16p11.2^{flx}* mice to express these transgenes in DR neurons lacking 16p11.2 and implanted an optic fibre above the DR (Extended Data Fig. 4a, b). As expected (Fig. 3k, l), eYFP mice displayed decreased sociability, whereas optogenetic activation of DR neurons rescued these sociability deficits (Extended Data Fig. 4c-f) with no changes in control behaviours (Extended Data Fig. 4g-i).

To determine whether 5-HT neuron activation specifically is sufficient to reverse the sociability deficits caused by 16p11.2 deletion, *Sert-cre:16p11.2^{flx}* mice were injected with DIO-ChR2 or DIO-eYFP. Activation of DR 5-HT neurons rescued the social behaviour deficits

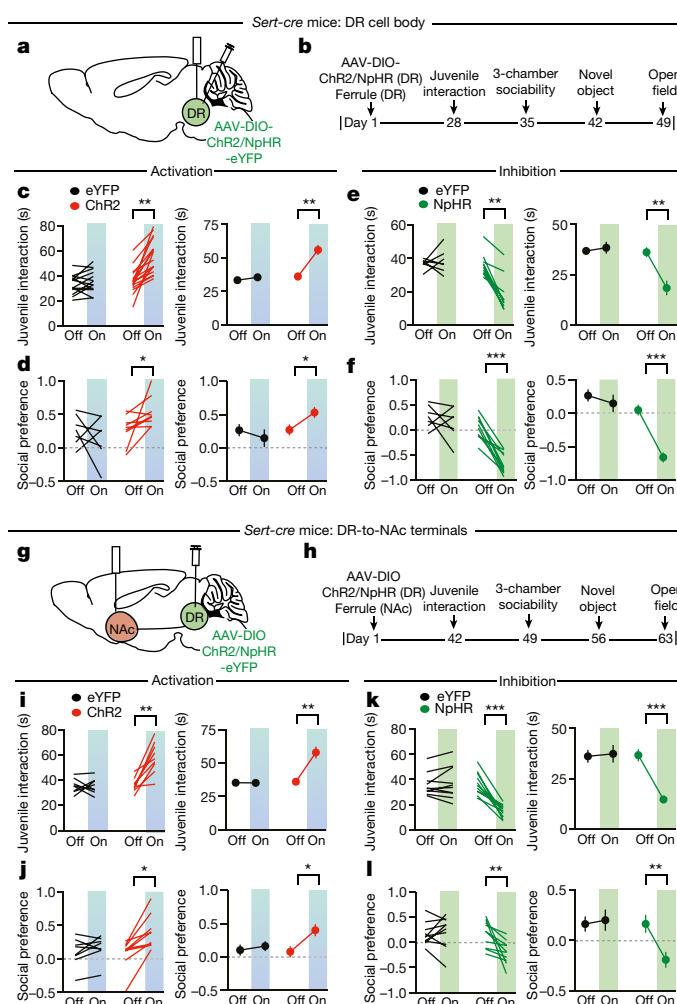


Fig. 2 | Bidirectional modulation of DR 5-HT neuron activity modifies sociability. **a**, Schematic of optogenetic manipulation. **b**, Timeline of experiments. **c**, **d**, Quantification of juvenile interaction (**c**: $F_{1,56} = 10.85$, $P < 0.01$; $n = 14-16$) and three-chamber sociability (**d**: $F_{1,28} = 4.597$, $P < 0.05$; $n = 7-9$) assays in *Sert-cre* mice expressing ChR2 or eYFP. **e**, **f**, Quantification of juvenile interaction (**e**: $F_{1,30} = 12.77$, $P < 0.01$; $n = 7-10$) and three-chamber sociability (**f**: $F_{1,30} = 11.98$, $P < 0.001$; $n = 7-10$) assays in *Sert-cre* mice expressing NpHR or eYFP. **g**, Schematic of optogenetic manipulation. **h**, Timeline of experiments. **i**, **j**, Quantification of juvenile interaction (**i**: $F_{1,30} = 15.78$, $P < 0.01$; $n = 8-9$) and three-chamber sociability (**j**: $F_{1,30} = 6.413$, $P < 0.05$; $n = 8-9$) in *Sert-cre* mice expressing ChR2 or eYFP. **k**, **l**, Quantification of juvenile interaction (**k**: $F_{1,38} = 16.29$, $P < 0.001$; $n = 10-11$) and three-chamber sociability (**l**: $F_{1,36} = 5.66$, $P < 0.01$; $n = 10$) assays in *Sert-cre* mice expressing NpHR or eYFP. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

in mice with 16p11.2 deletion from 5-HT neurons (Fig. 5a, b; Extended Data Fig. 5a, b) while this manipulation again had no effect on the novel object interaction assay, locomotion or anxiety-related behaviours (Extended Data Fig. 5c–e). To determine whether enhancing 5-HT release specifically in the NAc would have the same effects, we repeated these manipulations but placed the optic fibres in the NAc. Activation of DR-to-NAc 5-HT terminals rescued the sociability deficits in mice with 16p11.2 deletion from 5-HT neurons (Fig. 5c, d; Extended Data Fig. 5f, g) and had no effect on control behaviours (Extended Data Fig. 5h–j).

Dopamine release in the NAc also enhances social interaction¹⁷ but is strongly reinforcing on its own^{27–30}. To examine whether 5-HT release in the NAc shares these features, we performed two assays: real-time

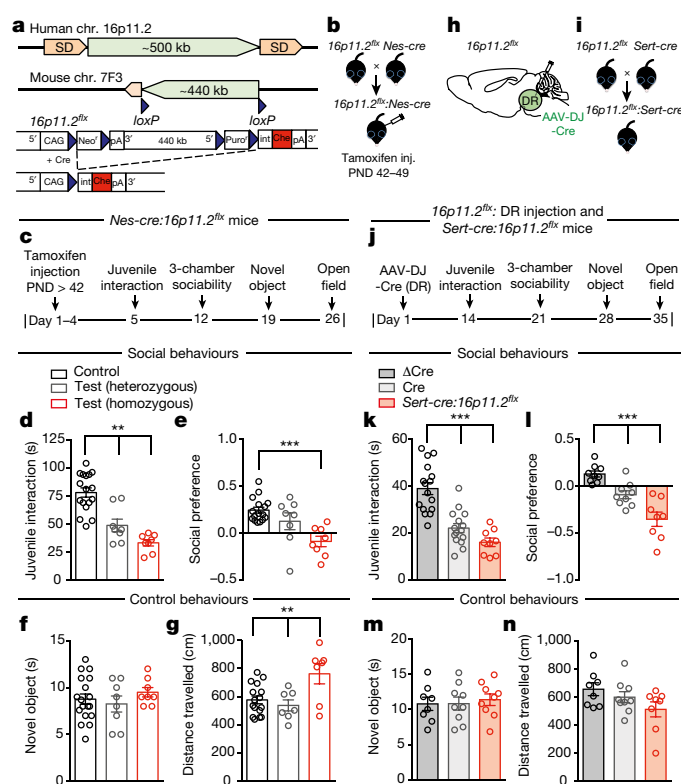


Fig. 3 | 16p11.2 deletion in DR and 5-HT neurons decreases sociability. **a**, Schematic of human chromosome 16p11.2 and deletion of syntenic region of mouse chromosome 7F3. **b**, Genetic crosses used to delete 16p11.2 from whole brain. PND, postnatal day. **c**, Timeline of experiments. **d**, **e**, Quantification of juvenile interaction (**d**: $F_{2,30} = 26.98$, $P < 0.01$; $n = 8-17$) and three-chamber sociability (**e**: $F_{2,29} = 10.03$, $P < 0.001$; $n = 8-16$) in control, heterozygous and homozygous mice with deletion of 16p11.2. **f**, **g**, Deletion of 16p11.2 does not alter the novel object interaction assay (**f**: $F_{2,31} = 0.6613$, $P = 0.5233$; $n = 8-18$), but homozygous 16p11.2 deletion increases locomotor activity (**g**: $F_{2,27} = 6.341$, $P < 0.01$; $n = 7-16$). **h**, Schematic of 16p11.2 deletion in DR. **i**, Genetic crosses to delete 16p11.2 from 5-HT neurons. **j**, Timeline of experiments. **k**, **l**, Quantification of juvenile interaction (**k**: $F_{2,37} = 26.72$, $P < 0.001$; $n = 9-16$) and three-chamber sociability (**l**: $F_{2,23} = 21.45$, $P < 0.001$; $n = 8-9$) assays in 16p11.2^{flx} mice expressing Δ Cre or Cre in DR and *Sert-cre*:16p11.2^{flx} mice. **m**, **n**, Deletion of 16p11.2 did not alter the novel object interaction assay (**m**: $F_{2,23} = 0.138$, $P = 0.8718$, $n = 8-9$) or the locomotion assay (**n**: $F_{2,22} = 2.371$, $P = 0.1168$, $n = 8-9$). Data are mean \pm s.e.m. ** $P < 0.01$; *** $P < 0.001$; one-way ANOVA with Tukey's multiple comparison post hoc test. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

conditioned place preference (CPP) and optogenetic intracranial self-stimulation. Optogenetic activation of DR-to-NAc 5-HT inputs did not elicit real-time CPP or nose poking for stimulation in the optogenetic intracranial self-stimulation protocol in either *Sert-cre* or *Sert-cre*:16p11.2^{flx} mice (Extended Data Fig. 6a–e). These results suggest that 5-HT release in the NAc, unlike dopamine release, is not acutely reinforcing.

To assess the specificity of 5-HT action in the NAc, we asked whether activation of 5-HT inputs to the NAc enhanced interactions with a non-social appetitive stimulus by performing a three-chamber test during which a high-fat food pellet was placed in one chamber. Control *Sert-cre* mice spent more time in the chamber containing the food pellet, confirming that it was appetitive, but activation of 5-HT inputs in the NAc did not increase preference for this food chamber (Extended Data Fig. 6f). Similar negative results were obtained in *Sert-cre*:16p11.2^{flx} mice (Extended Data Fig. 6g). To address whether 5-HT release in brain regions other than the NAc influences sociability, we activated 5-HT terminals in the dorsal striatum, a site that subserves

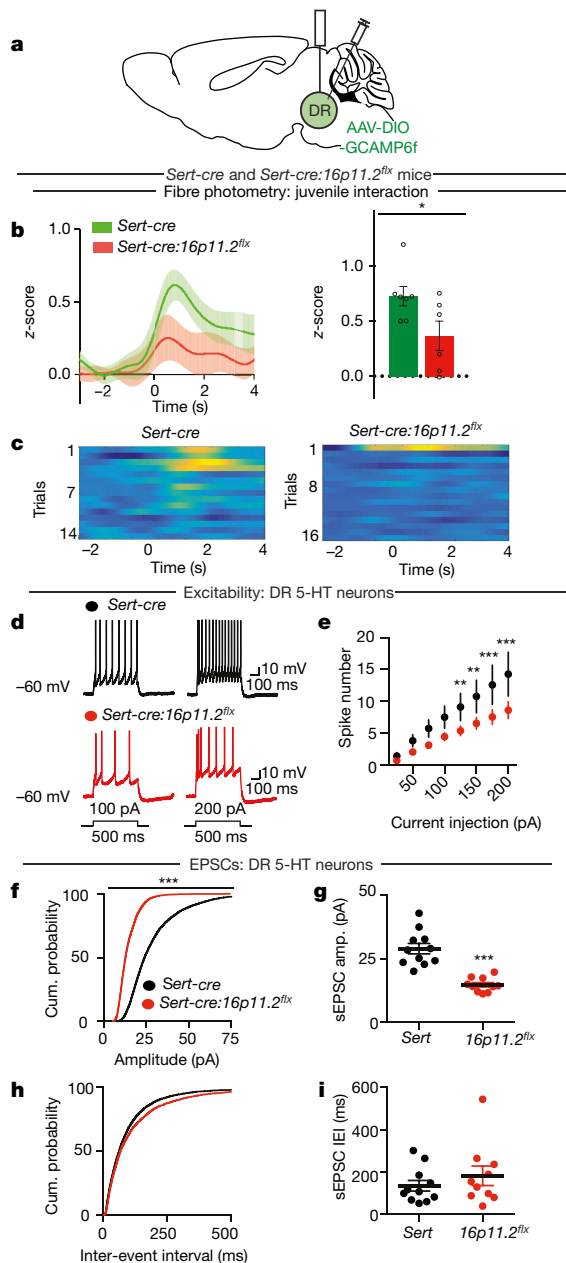


Fig. 4 | 16p11.2 deletion in DR 5-HT neurons decreases their activity. **a**, Schematic of experimental set-up. **b**, Left, time course of average GCaMP6f transient z-scores event-locked to social interaction. Right, quantification of average peak z-score during social interaction ($t_{11} = 2.321$, $P < 0.05$). **c**, Representative heat map of z-score changes over all trials from single mice. **d**, Sample traces of spiking in 5-HT DR neurons. **e**, Quantification of spiking ($F_{7,56} = 2.305$, $P < 0.05$). **f**, Summary of cumulative probability of spontaneous excitatory postsynaptic current (sEPSC) amplitudes. **g**, Mean sEPSC amplitude changes ($t_{19} = 0.069$, $P < 0.001$, $n = 10-11$). **h**, Summary of cumulative probability of sEPSC frequency. **i**, Mean sEPSC frequency changes. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; unpaired t -test (**b**, **g**, **i**), repeated measures two-way ANOVA with Sidak's multiple comparison post hoc test (**d**), or Kolmogorov-Smirnov test (**f**, **h**). The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

different functions from the NAc³¹⁻³⁴. Activation of DR 5-HT terminals in the dorsal striatum had no effect in the sociability assays (Extended Data Fig. 7a-f) or in the novel object interaction assay or locomotion in the open field (Extended Data Fig. 7g, h). However, 5-HT terminal activation in dorsal striatum did cause a decrease in the open field centre time in *Sert-cre* mice (Extended Data Fig. 7i), demonstrating the stimulation was effective.

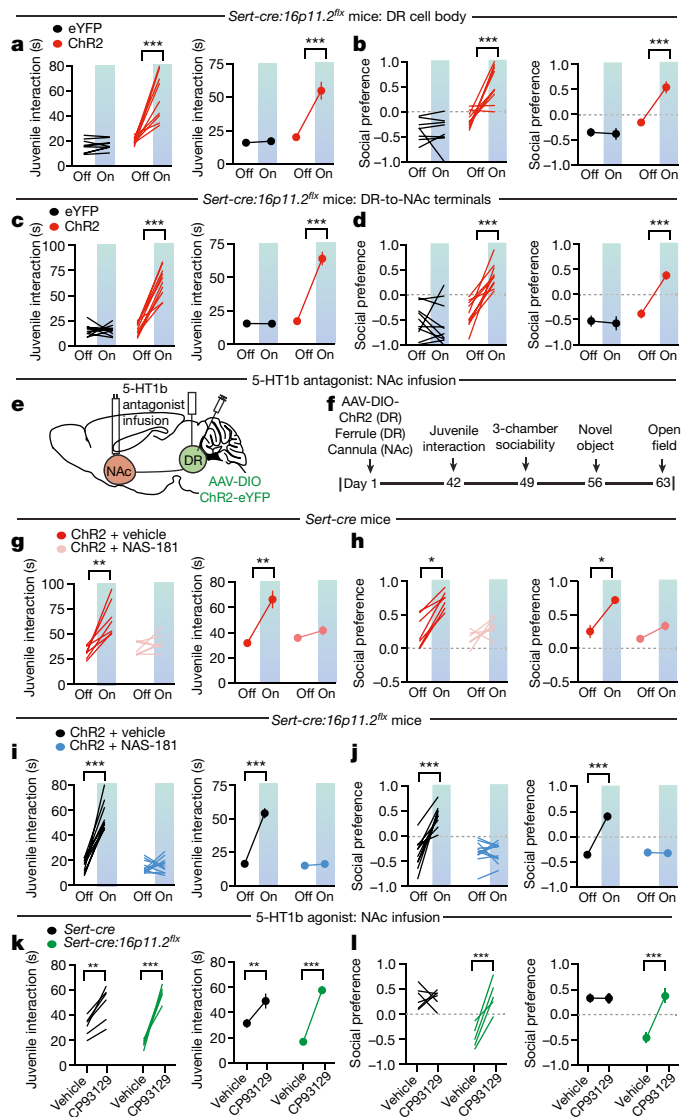


Fig. 5 | Rescue of social deficits in 16p11.2 deletion mice by 5-HT activity in the NAc. **a**, **b**, Quantification of juvenile interaction (**a**: $F_{1,32} = 25.15$, $P < 0.001$; $n = 9$) and three-chamber sociability (**b**: $F_{1,32} = 16.63$, $P < 0.001$; $n = 8-10$) assays in *Sert-cre:16p11.2flx* mice expressing ChR2 or eYFP receiving soma stimulation. **c**, **d**, Quantification of juvenile interaction (**c**: $F_{1,36} = 69.84$, $P < 0.001$; $n = 10$) and three-chamber sociability (**d**: $F_{1,36} = 16.46$, $P < 0.001$; $n = 10$) assays in *Sert-cre:16p11.2flx* mice expressing ChR2 or eYFP receiving DR-to-NAc terminal stimulation. **e**, Schematic of experimental set-up. **f**, Timeline of experiments. **g**, **h**, Quantification of juvenile interaction (**g**: $F_{1,24} = 12.6$, $P < 0.01$; $n = 7$) and three-chamber sociability (**h**: $F_{1,22} = 4.322$, $P < 0.05$; $n = 7$) in *Sert-cre* mice expressing ChR2 in DR with either vehicle or NAS-181 infused into NAc. **i**, **j**, Quantification of juvenile interaction (**i**: $F_{1,40} = 29.24$, $P < 0.001$; $n = 11$) and three-chamber sociability (**j**: $F_{1,38} = 29.06$, $P < 0.001$; $n = 10-11$) in *Sert-cre:16p11.2flx* mice expressing ChR2 in DR with either vehicle or NAS-181 infused into NAc. **k**, **l**, Quantification of juvenile interaction (**k**: $F_{1,9} = 19.03$, $P < 0.001$; $n = 10-11$) and three-chamber sociability (**l**: $F_{1,9} = 18.4$, $P < 0.01$; $n = 5-6$) in *Sert-cre* and *Sert-cre:16p11.2flx* mice with either vehicle or CP93129 infused into NAc. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; two-way ANOVA with Sidak's multiple comparison post hoc test. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

A question with therapeutic implications is whether the enhancement of sociability caused by the activation of DR-to-NAc 5-HT terminals outlasts the stimulation. To address this topic, we activated DR-to-NAc 5-HT terminals in *Sert-cre* and *Sert-cre:16p11.2flx* mice on

several days. Although stimulation on day 1 caused robust increases in sociability (Extended Data Fig. 8a), there was no lasting effect when these mice were assayed the next day even though they again showed an acute enhancement of sociability during stimulation (Extended Data Fig. 8b). To assess whether more prolonged terminal stimulation of DR-to-NAc 5-HT neurons could elicit effects beyond the stimulation period, we performed stimulation assays four times with 2 h between sessions on days 3 and 4: mice thereby received 10 bouts of DR-to-NAc 5-HT terminal stimulation over 4 days. Mice consistently showed enhanced sociability during each stimulation bout, but no carry over effects were observed (Extended Data Fig. 8c–j).

Role of NAc 5-HT1b receptors in enhanced sociability

Blockade of 5-HT1b receptors in the NAc abolishes the reinforcing properties of social interaction in a CPP assay³. To test whether the effects of DR 5-HT neuron activation on sociability also require NAc 5-HT1b receptors, we infused a 5-HT1b receptor antagonist (NAS-181) into the NAc before optogenetic behavioural assays in *Sert-cre* mice (Fig. 5e, f). As expected (Fig. 2c, d), the activation of DR 5-HT neurons caused increases in sociability, which were blocked by NAS-181 infusions into the NAc (Fig. 5g, h; Extended Data Fig. 9a, b). Identical NAc infusions of NAS-181 had no effects on sociability assays in control *Sert-cre* mice expressing eYFP (Extended Data Fig. 9c–f) nor on control behaviours in any of the cohorts of mice (Extended Data Figs. 9g–i, 10a–c). We next examined whether the rescue of sociability deficits by DR 5-HT neuron stimulation in mice with 16p11.2 deletion was also dependent on NAc 5-HT1b receptors. The increase in sociability elicited by DR 5-HT neuron activation was again observed in *Sert-cre:16p11.2^{flx}* mice that received NAc infusions of vehicle, and these behavioural effects were prevented by infusions of NAS-181 (Fig. 5i, j; Extended Data Fig. 10d, e) with no effects on control behaviours in either cohort of mice (Extended Data Fig. 10f–h).

In final experiments, we asked whether pharmacological activation of NAc 5-HT1b receptors was sufficient to rescue the social deficits in 16p11.2 deletion mice. Both *Sert-cre* and *Sert-cre:16p11.2^{flx}* mice exhibited an increase in sociability during the juvenile interaction assay after direct infusion of the 5-HT1b receptor agonist CP93129 into the NAc (Fig. 5k). During the three-chamber social preference assay, NAc infusion of CP93129 robustly rescued the behavioural deficits in *Sert-cre:16p11.2^{flx}* mice while having no detectable effect in control *Sert-cre* mice (Fig. 5l, Extended Data Fig. 11a–f). Control behaviours in both cohorts of mice were unaffected by this drug treatment (Extended Data Fig. 11g–i).

Concluding remarks

Our results demonstrate that stimulating 5-HT release in the NAc promotes sociability, effects that require activation of NAc 5-HT1b receptors. Inhibition of DR 5-HT neurons or their terminals in the NAc reduced social interactions, suggesting that 5-HT action in the NAc is necessary for normal levels of sociability. Consistent with this conclusion, increases in DR 5-HT neuron activity occur during non-aggressive social interactions, including mating, although increases may also occur during sucrose or food intake⁹. Stimulating 5-HT release in the NAc did not elicit acute reinforcement or changes in a variety of control behaviours. These behavioural effects of 5-HT in the NAc are markedly different from the acute reinforcing properties of the release of dopamine in the NAc^{27–30}, suggesting critical differences in the NAc circuitry modulation by which these major neuromodulators mediate their behavioural effects.

The findings that genetic deletion of 16p11.2 specifically from 5-HT neurons caused sociability deficits accompanied by decreases in DR 5-HT neuron activity during social interactions as well as in vitro electrophysiological assays provide further evidence for the importance of DR 5-HT neuron activity in prosocial behaviours. However, further work is necessary to determine how accurately the postnatal, spatially restricted homozygous deletion of 16p11.2 in mice mimics the human heterozygous deletion syndrome and whether other autism models

will express deficits in DR 5-HT neuron functioning. The rescue of sociability deficits by DR 5-HT neuron activation raises the question of whether drugs that influence 5-HT levels are beneficial in treating ASD. Serotonin reuptake inhibitors treat anxiety and obsessive-compulsive behaviours associated with ASD with variable efficacy, but there is little evidence that they ameliorate social deficits³⁵. Our results demonstrate that direct activation of NAc 5-HT1b receptors or increased 5-HT release in the NAc is sufficient to ameliorate social deficits. Unlike the optogenetic inhibition of 5-HT neuron activity (Fig. 2), however, NAc infusion of the 5-HT1b receptor antagonist did not impair baseline sociability (Fig. 5). These results suggest that the influence of 5-HT in the NAc on sociability probably involves other subtypes of 5-HT receptor.

Another potential pharmacological intervention for promoting sociability is MDMA (3,4-methylenedioxymethamphetamine), which robustly releases 5-HT in an activity-independent manner and is being tested for its therapeutic utility³⁶. Optogenetic stimulation of DR 5-HT neurons is more likely to mimic the effects of MDMA on 5-HT levels in target structures, than serotonin reuptake inhibitors. Thus, like dopamine³⁷, the behavioural effects of drugs that influence 5-HT uptake or release probably depend on the specific influence of the drug on 5-HT levels and kinetics.

The DR is a heterogeneous structure that contains several cell types. Previous studies suggest that manipulations of the activity of these cells have complex effects on motivated behaviours^{9,38–45}. Assuming that the large body of work on dopamine modulation of target circuits and behaviours is an appropriate comparator^{27–30}, 5-HT modulation of circuits and behaviours will be equally complex and depend on its specific anatomical targets. Our findings suggest that detailed exploration of 5-HT function using modern circuit neuroscience tools will not only advance our understanding of the adaptive role of this neuromodulatory system but will also provide insights that will prove valuable for development of mechanistically novel therapies for the treatment of prevalent neuropsychiatric disorders such as ASD.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0416-4>.

Received: 22 June 2017; Accepted: 27 June 2018;

Published online 8 August 2018.

- Christensen, D. L. et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 Sites, United States, 2012. *MMWR Surveill. Summ.* **65**, 1–23 (2016).
- Chevallier, C., Kohls, G., Troiani, V., Brodtkin, E. S. & Schultz, R. T. The social motivation theory of autism. *Trends Cogn. Sci.* **16**, 231–239 (2012).
- Dölen, G., Darvishzadeh, A., Huang, K. W. & Malenka, R. C. Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. *Nature* **501**, 179–184 (2013).
- Brown, S.-L. & Praag, H. M. v. *The Role of Serotonin in Psychiatric Disorders* (Brunner/Mazel, New York, 1991).
- Charney, D. S., Sklar, P. B., Buxbaum, J. D. & Nestler, E. J. *Charney & Nestler's Neurobiology of Mental Illness* 5th edn (Oxford Univ. Press, 2018).
- Furay, A. R., McDewitt, R. A., Miczek, K. A. & Neumaier, J. F. 5-HT1B mRNA expression after chronic social stress. *Behav. Brain Res.* **224**, 350–357 (2011).
- Kane, M. J. et al. Mice genetically depleted of brain serotonin display social impairments, communication deficits and repetitive behaviors: possible relevance to autism. *PLoS One* **7**, e48975 (2012).
- Challis, C. et al. Raphe GABAergic neurons mediate the acquisition of avoidance after social defeat. *J. Neurosci.* **33**, 13978–13988, 13988a (2013).
- Li, Y. et al. Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nat. Commun.* **7**, 10503 (2016).
- Muller, C. L., Anacker, A. M. J. & Veenstra-VanderWeele, J. The serotonin system in autism spectrum disorder: From biomarker to animal models. *Neuroscience* **321**, 24–41 (2016).
- Schain, R. J. & Freedman, D. X. Studies on 5-hydroxyindole metabolism in autistic and other mentally retarded children. *J. Pediatr.* **58**, 315–320 (1961).
- Weiss, L. A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
- Kumar, R. A. et al. Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).

14. Sanders, S. J. et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
15. Christoffel, D. J. et al. IKB kinase regulates social defeat stress-induced synaptic and behavioral plasticity. *J. Neurosci.* **31**, 314–321 (2011).
16. Walsh, J. J. et al. Stress and CRF gate neural activation of BDNF in the mesolimbic reward pathway. *Nat. Neurosci.* **17**, 27–29 (2014).
17. Gunaydin, L. A. et al. Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535–1551 (2014).
18. Francis, T. C. et al. Nucleus accumbens medium spiny neuron subtypes mediate depression-related outcomes to social defeat stress. *Biol. Psychiatry* **77**, 212–222 (2015).
19. Wallace, D. L. et al. CREB regulation of nucleus accumbens excitability mediates social isolation-induced behavioral deficits. *Nat. Neurosci.* **12**, 200–209 (2009).
20. Luo, M., Zhou, J. & Liu, Z. Reward processing by the dorsal raphe nucleus: 5-HT and beyond. *Learn. Mem.* **22**, 452–460 (2015).
21. Gong, S. et al. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J. Neurosci.* **27**, 9817–9823 (2007).
22. Veenstra-VanderWeele, J. et al. Autism gene variant causes hyperserotonemia, serotonin receptor hypersensitivity, social impairment and repetitive behavior. *Proc. Natl Acad. Sci. USA* **109**, 5469–5474 (2012).
23. Portmann, T. et al. Behavioral abnormalities and circuit defects in the basal ganglia of a mouse model of 16p11.2 deletion syndrome. *Cell Reports* **7**, 1077–1092 (2014).
24. Burns, K. A. et al. Nestin-CreER mice reveal DNA synthesis by nonapoptotic neurons following cerebral ischemia hypoxia. *Cereb. Cortex* **17**, 2585–2592 (2007).
25. Horev, G. et al. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc. Natl Acad. Sci. USA* **108**, 17076–17081 (2011).
26. Grimm, D. et al. *In vitro* and *in vivo* gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82**, 5887–5911 (2008).
27. Tsai, H. C. et al. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080–1084 (2009).
28. Stuber, G. D., Britt, J. P. & Bonci, A. Optogenetic modulation of neural circuits that underlie reward seeking. *Biol. Psychiatry* **71**, 1061–1067 (2012).
29. Steinberg, E. E. & Janak, P. H. Establishing causality for dopamine in neural function and behavior with optogenetics. *Brain Res.* **1511**, 46–64 (2013).
30. Lammel, S., Lim, B. K. & Malenka, R. C. Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* **76 Pt B**, 351–359 (2014).
31. Steinbusch, H. W., van der Kooy, D., Verhofstad, A. A. & Pellegrino, A. Serotonergic and non-serotonergic projections from the nucleus raphe dorsalis to the caudate-putamen complex in the rat, studied by a combined immunofluorescence and fluorescent retrograde axonal labeling technique. *Neurosci. Lett.* **19**, 137–142 (1980).
32. Steinbusch, H. W. Distribution of serotonin-immunoreactivity in the central nervous system of the rat-cell bodies and terminals. *Neuroscience* **6**, 557–618 (1981).
33. Hornung, J. P. The human raphe nuclei and the serotonergic system. *J. Chem. Neuroanat.* **26**, 331–343 (2003).
34. Michelsen, K. A., Prickaerts, J. & Steinbusch, H. W. The dorsal raphe nucleus and serotonin: implications for neuroplasticity linked to major depression and Alzheimer's disease. *Prog. Brain Res.* **172**, 233–264 (2008).
35. Politte, L. C., Henry, C. A. & McDougle, C. J. Psychopharmacological interventions in autism spectrum disorder. *Harv. Rev. Psychiatry* **22**, 76–92 (2014).
36. Heifets, B. D. & Malenka, R. C. MDMA as a probe and treatment for social behaviors. *Cell* **166**, 269–272 (2016).
37. Volkow, N. D., Fowler, J. S., Wang, G. J. & Swanson, J. M. Dopamine in drug abuse and addiction: results from imaging studies and treatment implications. *Mol. Psychiatry* **9**, 557–569 (2004).
38. Qi, J. et al. A glutamatergic reward input from the dorsal raphe to ventral tegmental area dopamine neurons. *Nat. Commun.* **5**, 5390 (2014).
39. Liu, Z. et al. Dorsal raphe neurons signal reward through 5-HT and glutamate. *Neuron* **81**, 1360–1374 (2014).
40. McDevitt, R. A. et al. Serotonergic versus nonserotonergic dorsal raphe projection neurons: differential participation in reward circuitry. *Cell Reports* **8**, 1857–1869 (2014).
41. Matthews, G. A. et al. Dorsal raphe dopamine neurons represent the experience of social isolation. *Cell* **164**, 617–631 (2016).
42. Warden, M. R. et al. A prefrontal cortex-brainstem neuronal projection that controls response to behavioural challenge. *Nature* **492**, 428–432 (2012).
43. Fonseca, M. S., Murakami, M. & Mainen, Z. F. Activation of dorsal raphe serotonergic neurons promotes waiting but is not reinforcing. *Curr. Biol.* **25**, 306–315 (2015).
44. Correia, P. A. et al. Transient inhibition and long-term facilitation of locomotion by phasic optogenetic activation of serotonin neurons. *eLife* **6**, e20975 (2017).
45. Marcinkiewicz, C. A. et al. Serotonin engages an anxiety and fear-promoting circuit in the extended amygdala. *Nature* **537**, 97–101 (2016).
46. Franklin, K. B. J. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* 4th edn (Academic, 2012).

Acknowledgements This study was supported by the Simons Foundation Autism Research Initiative (award 305112 to R.C.M.) and NIMH (F32 MH103949 to J.J.W.).

Reviewer information Nature thanks G. Feng, M. Lobo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.J.W. performed the majority of experiments. D.J.C. performed the electrophysiology experiments. B.D.H. performed the fibre photometry experiments. G.A.B.-D., A.S. and L.W.H. assisted in surgeries and behavioural assays. J.J.W. and R.C.M. designed the experiments, interpreted results and wrote the paper, which was edited by all authors.

Competing interests R.C.M. and K.D. are cofounders and on the scientific advisory board of Circuit Therapeutics, Inc.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0416-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0416-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.C.M. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Experimental subjects. Male 7–9-week-old C57BL/6 mice (Jackson Laboratory), Tg(Slc6a4-cre)ET33Gsat (*Sert-cre*, Jackson Laboratory)²¹, C57BL/6-Tg(Nes-cre/Esr1*)1Kuan/J (*Nes-creER*, Jackson Laboratory)²⁴, and B6N.129P2(Cg)-Igs13^{tm1Dolm}Igs14^{tm1Dolm}/J (*16p11.2^{flx}*, gift from R. Dolmetsch, Jackson Laboratory)²³ mice were used as experimental subjects alone or following the crosses described in the text. All wild-type and *Sert-cre* mice were on C57BL/6 backgrounds. *16p11.2^{flx}* and *16p11.2^{flx};Nes-cre* mice were on CD1 backgrounds. *Sert-cre:16p11.2^{flx}* mice were on a mixed background of C57BL/6 and CD1. In all experiments, controls were littermates on the exact same background. Juvenile mice used for the juvenile interaction assay and three-chamber sociability assay were male, conspecific and 3–5 weeks of age. Mice were housed on a 12-h light/dark cycle with food and water ad libitum. All procedures complied with the animal care standards set forth by the National Institute of Health and were approved by Stanford University's Administrative Panel on Laboratory Animal Care and Administrative Panel of Biosafety. No statistical methods were used to predetermine sample size. All experiments were conducted in a blinded manner such that assays were conducted and analysed without knowledge of the specific manipulation being performed and with animals being randomized by cage before surgery and behavioural experiments.

Viral vectors and stereotaxic surgeries for optogenetic methods. AAVs used in this study that were purchased from the University of North Carolina Viral Core included: AAV-ChR2-eYFP, AAV-eYFP, AAV-DIO-ChR2-eYFP, AAV-DIO-NpHR-eYFP, AAV-DIO-eYFP, AAV-DJ-Cre, AAV-DR-ΔCre and AAV-DJ-DIO-GCaMP6f were purchased from the Stanford Neuroscience Gene Vector and Virus Core. For surgeries, mice (7–9 weeks of age) were anaesthetized with a mixture of ketamine (100 mg kg⁻¹) and xylazine (10 mg kg⁻¹), positioned in a small-animal stereotaxic instrument (Kopf Instruments) and the skull surface was exposed. Thirty-three gauge syringe needles (Hamilton) were used to unilaterally infuse 0.3 μl of virus into the DR (bregma coordinates: anteroposterior, -4.36; mediolateral, 0; dorsoventral, -3.1) at a rate of 0.1 μl min⁻¹. Needles were removed 5 min after infusions were complete.

For optogenetic behavioural experiments, optic fibres (ferrules) were implanted above the DR (bregma coordinates: anteroposterior, -4.36; mediolateral, 0; dorsoventral, -3.0) for somatic stimulation or above the NAc bilaterally (bregma coordinates: anteroposterior, +1.5; mediolateral, ±0.75; dorsoventral, -3.9) or dorsal striatum bilaterally (bregma coordinates: anteroposterior, +0.5; mediolateral, ±1.6; dorsoventral, -2.7) for terminal stimulation. Ferrules were made in-house using 1.25 mm diameter multimode ceramic ferrules (ThorLab), 200 μm fibre optic cable with numerical aperture (NA) 0.39 (ThorLabSA) and blue dye epoxy (Fibre Instrument Sales). Ferrules were secured to the skull using miniature screws (thread size 00–90 × 1/16, Antrin Miniature Specialties) and light-cured dental adhesive cement (Geristore A&B paste, DenMat).

For drug infusions, a 26-gauge guide cannula, 3.6 mm in length from the cannula base, was implanted bilaterally into the NAc (bregma coordinates: anteroposterior, +1.5; mediolateral, ±0.75; dorsoventral, -3.9).

Optogenetic stimulation. For optogenetic photostimulation, ferrules were connected to a 473 nm laser diode (OEM Laser Systems) through a FC/PC adaptor and a fibre optic rotary joint (Doric Lenses). Laser output was controlled using a Master-8 pulse stimulator (A.M.P.I.), which delivered 5 ms light pulses at 20 Hz^{9,39,40,47,48}. Light output through the optical fibres was adjusted to ~5 mW (somatic) or ~15 mW (terminals) using a digital power meter console (ThorLabs). For activation of NpHR3.0, the optical fibre was connected to a 532 nm laser diode (Shanghai Dream Lasers Technology Co, Ltd) via a FC/PC adaptor and a fibre optic rotary joint (Doric Lenses). Laser output was again controlled using a Master-8 pulse stimulator (A.M.P.I.) and adjusted to ~10 mW. Mice received cycles of 8 s light on and 2 s light off.

Microinjection. One hour before behavioural experiments, mice received an intra-NAc infusion of 5-HT1b antagonist (NAS-181, 1.5 μg) or 5-HT1b agonist (CP93129, 0.5 μg) (Tocris Biosciences) or vehicle. Drugs were infused through an injector cannula using a microinfusion pump (Harvard Apparatus) at a continuous rate of 0.1 μl per min to a total volume of 0.3 μl per hemisphere. Injector cannulae were removed 2 min after infusions were complete, and mice were allowed to sit undisturbed for 1 h before behavioural tests.

Juvenile interaction assay. Juvenile interaction was performed in the home cage of the test animal as previously described¹⁷. Cagemates were temporarily moved to a holding container and the test mouse was habituated for 1 min. For optogenetic experiments, the fibre optic patch-cord was connected during the 1 min habituation period. After habituation, a novel conspecific juvenile mouse (3–5-week-old males) was placed into the home cage for 2 min of free interaction, with the laser on for the duration of the session during stimulation rounds. All sessions were video recorded with a camera located above the home cage and analysed manually following the behaviour. Interaction time was defined as those times during which the test mouse was actively exploring the juvenile mouse as defined by active

pursuit, sniffing any region (including the snout, body, and anogenital area) as well as grooming. These individual social behaviours were not assayed independently. Each test mouse underwent two rounds of the juvenile interaction assay, separated by 1 h, with a novel juvenile introduced during each session. Cohorts of mice were counterbalanced for the order of providing optogenetic stimulation versus no stimulation. All experiments and analyses were performed blinded without knowledge of the manipulation to which the subject had been subjected and the genotype of the subject.

Three-chamber sociability assay. A three-chamber sociability assay was performed in an arena with three separate chambers as previously described⁴⁹. On day one, the test mice were habituated to the arena, with two empty wire mesh cups placed in the two outer chambers for 5 min. Male, conspecific juvenile mice also habituated to the mesh cups for 5 min following test mice habituation. On day two, the test mouse was placed in the centre chamber and a conspecific juvenile (3–5-week-old males) was placed into one of the wire mesh cups. The tops of the mesh cups were covered so as to prevent mice from crawling on top. They were immobilized on the floor of the chambers and had mesh with square holes that were 0.8 × 0.8 cm. The test mice were placed in the centre chamber for 2 min. The barriers were then raised and the test mouse was allowed to explore freely for a 20 min session. For animals receiving optogenetic stimulation, mice had 5 min epochs of the laser being off or on, which were counterbalanced across mice. The placement of juvenile mice in the chamber was also counterbalanced across sessions. For the three-chamber assay using a high-fat food pellet, the pellet was placed under the wire mesh cup instead of a juvenile. Location of mice was assayed automatically using a video tracking system (BIOBEHAVE). Sociability was calculated as: ((time in juvenile side - time in empty side)/(time in juvenile side + time in empty side)). These analyses are shown in Figs. 1–3 and 5. The actual times spent in each chamber during each 5 min epoch in all mouse cohorts are shown in Extended Data Figs. 1–11, which also illustrate the time spent within 3 cm of the mesh wire cups containing the juvenile (defined as proximity time).

Novel object interaction assay. The novel object interaction assay was performed in the exact same manner as the juvenile interaction assay, with either a toy mouse or a plastic block placed into the animal's home cage. The total time of investigation was again 2 min.

Open field test. To assay the effects of the different manipulations on locomotor activity, an open field test was conducted. Test mice were placed in an open field arena (40 × 40 cm) and allowed to move freely for an 18-min session. For animals receiving optogenetic stimulation, mice had 3 min epochs of laser being off or on, which were counterbalanced across mice. Time spent in the centre (25 × 25 cm) of the arena was also assayed as a measure of anxiety-related behaviour. Total distance travelled and centre time was assayed automatically using the video tracking system (BIOBEHAVE) and compared between light on and light off epochs.

Real-time CPP. The real-time CPP protocol was conducted as described previously⁵⁰ in a rectangular Plexiglas cage with three chambers separated by removable Plexiglass walls. The left and right chambers each measured 28 × 24 cm and had distinct wall patterns (black and white stripes versus black and white squares) and flooring (smooth versus rough plastic floors). The centre chamber measured 11.5 × 24 cm with no wall patterns and a smooth clear floor. Subjects were placed in the centre compartment for 2 min at which point the barriers were lifted and the subject mouse was allowed to freely explore the entire apparatus for 15 min during which it was photostimulated (20 Hz, 5 ms pulses) whenever it entered the designated chamber, which was alternated between each testing session. Video tracking software (BIOBEHAVE) recorded all animal movements and automatically analysed time spent and distance moved in each chamber. Preference was calculated by measuring total time in each chamber (stimulated, non-stimulated and centre).

Optical intracranial self-stimulation. Sixty-minute behavioural sessions were conducted in conditioning chambers (Med Associates Inc.) contained within sound-attenuating cubicles. Session start was indicated to the mouse by the illumination of a chamber light and the onset of low-volume white noise (65 dB) to mask external sounds. Two nosepoke ports, designated 'active' and 'inactive', were positioned on the left chamber wall. A response at the active nosepoke port resulted in optical stimulation (60 pulses, 5 ms duration, 20 Hz, 473 nm) on a fixed-ratio 1 schedule, with the exception that a new stimulation train could not be earned until any ongoing train had finished. Responses at the inactive nosepoke port were recorded but had no consequence. During the first training session, both nosepoke ports were baited with a crushed treat to facilitate initial investigation.

Fibre photometry. AAV-DJ-DIO-GCaMP6f was infused into the DR at the same stereotaxic coordinates noted previously, and a fibreoptic implant was advanced and secured at the same location. After allowing 3–4 weeks for viral expression, mice were first habituated to the fibre photometry apparatus for 30 min, and then tested on a subsequent day. Behaviour testing consisted of the juvenile interaction assay, as described above, with continuous video and fibre photometry acquisition. Fibre photometry data was acquired with Synapse software controlling an RZ5P lock-in amplifier (Tucker-Davis Technologies). GCaMP6f excitation was achieved

with a 473 nm LED (Doric), emission was measured with a femtowatt photoreceiver (2151; Newport), and signal was digitized as 6 kHz. All optical signals were band-pass filtered with a Fluorescence MiniCube FMC4 (Doric).

Signal processing was performed with Matlab (Mathworks, Inc.). In brief, signals were debleached by fitting with a mono- or bi-exponential decay function, and the resulting fluorescence trace was z-scored. Video was manually analysed by a genotype-blinded observer, who determined time of adult-to-juvenile contact. Peristimulus time histograms were constructed by taking the average of 7-s non-overlapping epochs of fluorescence consisting of 3 s before, and 4 s after contact time, which is defined as time = 0. Before averaging, each epoch was offset such that the z-score averaged from −3 to −1 s equalled 0. Peak z-scored fluorescence was determined for each peristimulus time histogram as the maximal z-score value between 0 and +4 s.

Electrophysiology. Mice were anaesthetized with isoflurane and coronal DR slices (250 μ m) were prepared after intracardiac perfusions with ice-cold artificial cerebrospinal fluid (aCSF) which contained (in mM): 128 NaCl, 3 KCl, 1.25 NaH₂PO₄, 10 D-glucose, 24 NaHCO₃, 2 CaCl₂ and 2 MgCl₂ (oxygenated with 95% O₂ and 5% CO₂, pH 7.4, 295–305 mOsm). Acute brain slices containing DR 5-HT neurons were generated using a microslicer (Leica VT1200S) in cold sucrose aCSF, which was derived by fully replacing NaCl with sucrose (254 mM) and saturated by 95% O₂ and 5% CO₂. Slices were maintained in holding chambers with aCSF for 1 h at 32 °C. Patch pipettes (3–5 M Ω) for whole-cell current-clamp and voltage-clamp recordings were pulled from borosilicate glass and filled with internal solution containing (in mM): 135 potassium gluconate, 10 HEPES, 4 KCl, 4 MgATP, and 0.3 NaGTP (pH 7.31, 287 mOsm).

Recordings from DR 5-HT neurons (identified visually by the presence of eYFP due to injection of DIO-eYFP into *Sert-cre* or *Sert-cre:16p11.2^{flx}* mice) were carried out in slices perfused with aCSF at 32 °C (flow rate = 2.5 ml min^{−1}). Recordings were made using a Multiclamp 700B amplifier. Signals were digitized at 8 kHz using an ITC-18 A/D converter (Instrutech Corporation) and filtered at 4 kHz. Data were acquired and analysed using Axograph-X (Axograph). For whole-cell recordings of spontaneous EPSCs, neurons were voltage clamped at −60 mV and series resistance (10–30 M Ω) was monitored throughout the recordings with

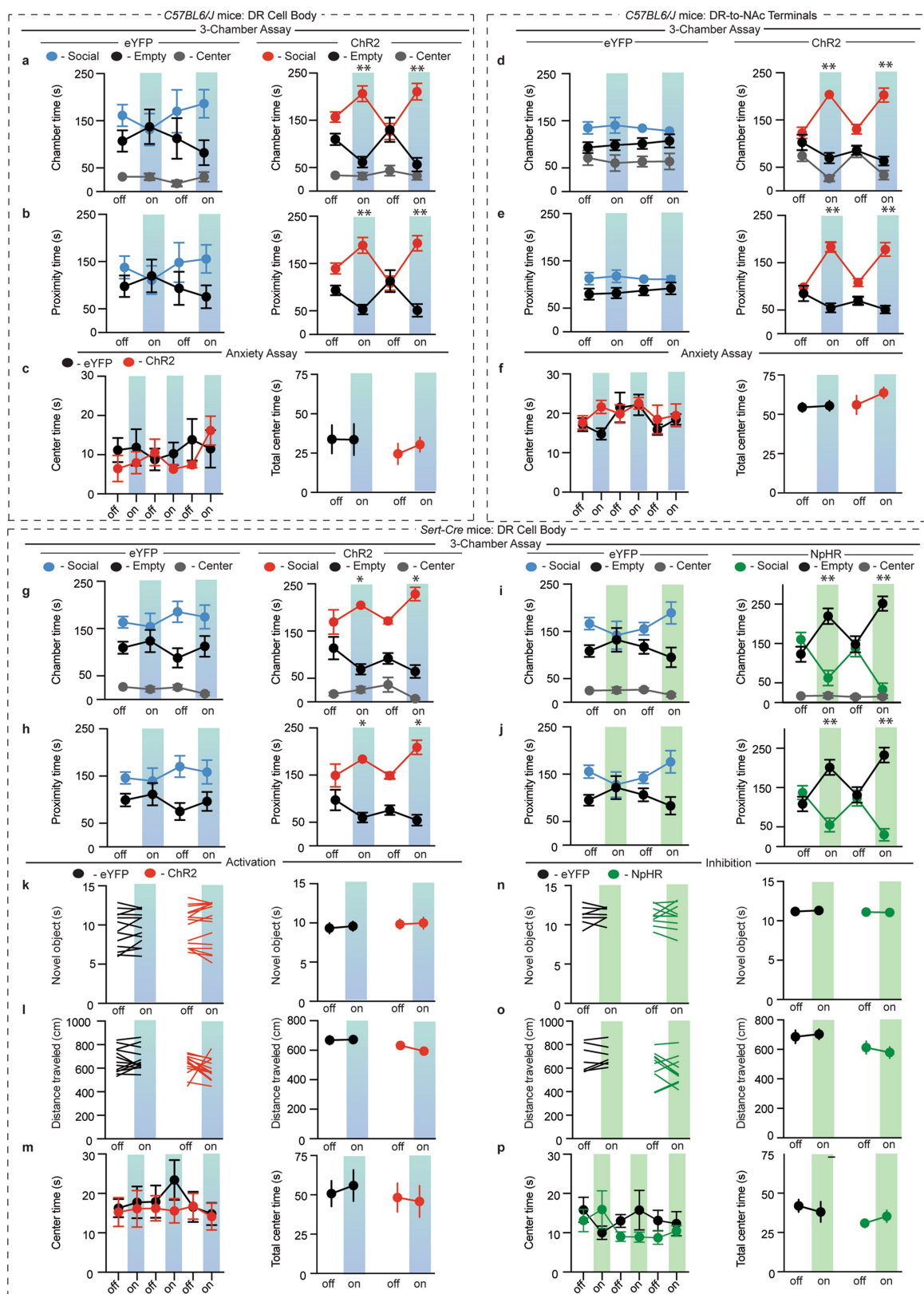
neurons discarded if resistance changed by >20%. At least 200 events per cell were acquired in 15 s blocks and detected using a threshold of 7 pA. Events were detected using Axograph event detection function and all events included in the final data analysis were verified by eye. To measure the intrinsic membrane properties of DR 5-HT neurons, whole-cell recordings were carried out in current-clamp mode at −60 mV and spikes were induced by incremental increases of current injection (each step increase was 50 pA; range 50–200 pA). Spike numbers were counted by eye. All data acquisition and analyses were performed blinded to the genotype of the cells from which recordings were made.

Blinding and statistics. As mentioned above, for all data acquisition and analyses in this study, investigators were blinded to the manipulation that the experimental subject had received and the genotype of the subject. Student's *t*-tests were used to compare two groups. Kolmogorov–Smirnov test was used for cumulative probability plots. One-way ANOVA with Tukey's post hoc test was used to compare multiple groups. Two-way ANOVA was used for analysis of multiple groups with Sidak's or Tukey's multiple comparison post hoc test, when appropriate. Statistical analyses were performed using Prism 6.0 (GraphPad Software). All data were tested and shown to exhibit normality and equal variances. All data are expressed as mean \pm s.e.m.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data are available from the corresponding author upon reasonable request.

47. Sharp, T., Bramwell, S. R., Clark, D. & Grahame-Smith, D. G. *In vivo* measurement of extracellular 5-hydroxytryptamine in hippocampus of the anaesthetized rat using microdialysis: changes in relation to 5-hydroxytryptaminergic neuronal activity. *J. Neurochem.* **53**, 234–240 (1989).
48. Hayashi, K., Nakao, K. & Nakamura, K. Appetitive and aversive information coding in the primate dorsal raphe nucleus. *J. Neurosci.* **35**, 6195–6208 (2015).
49. Kaidanovich-Beilin, O., Lipina, T., Vukobradovic, I., Roder, J. & Woodgett, J. R. Assessment of social interaction behaviors. *J. Vis. Exp.* **48**, 2473 (2011).
50. Lammel, S. et al. Input-specific control of reward and aversion in the ventral tegmental area. *Nature* **491**, 212–217 (2012).

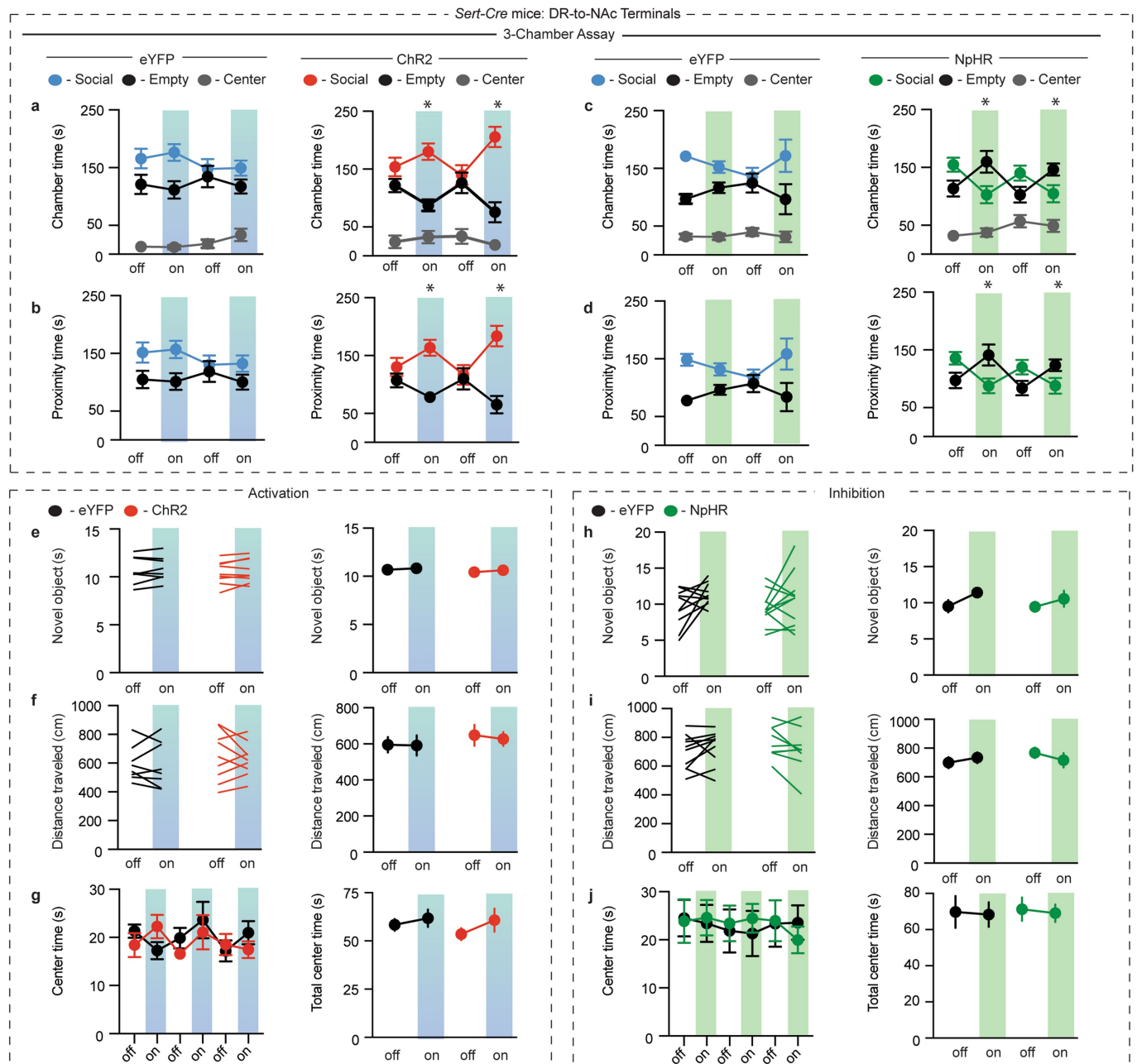


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Activation of DR neurons or their NAc projections increases sociability, and bidirectional modulation of DR 5-HT neuron activity bidirectionally modifies sociability.

a, b, Quantification of chamber time (**a**: eYFP, $F_{6,45} = 0.6823$, $P = 0.6647$, $n = 6$; ChR2, $F_{6,72} = 17.21$, $P < 0.01$; $n = 9$) and proximity time (**b**: eYFP, $F_{3,30} = 0.7517$, $P = 0.5300$, $n = 6$; ChR2, $F_{3,48} = 32.96$, $P < 0.01$, $n = 9$) in the three-chamber assay. **c**, Quantification of centre time in the locomotion assay (**c**: $F_{5,65} = 1.263$, $P = 0.2908$, $n = 6-9$). **d, e**, Quantification of chamber time (**d**: eYFP, $F_{6,54} = 0.2602$, $P = 0.9529$, $n = 7$; ChR2, $F_{6,72} = 19.36$, $P < 0.01$, $n = 9$) and proximity time (**e**: eYFP, $F_{3,36} = 0.3456$, $P = 0.7925$, $n = 7$; ChR2, $F_{3,48} = 26.44$, $P < 0.01$, $n = 9$) in the three-chamber assay. **f**, Quantification of centre time in the locomotion assay ($F_{5,40} = 0.7786$, $P = 0.5710$, $n = 5$). **g, h**, Quantification of chamber time (**g**: eYFP, $F_{6,54} = 0.7293$, $P = 0.6280$, $n = 7$; ChR2, $F_{6,72} = 3.812$, $P < 0.05$, $n = 9$) and proximity time (**h**: eYFP, $F_{3,36} = 0.9256$, $P = 0.4383$, $n = 7$; ChR2, $F_{3,48} = 4.844$, $P < 0.05$, $n = 9$) in the three-

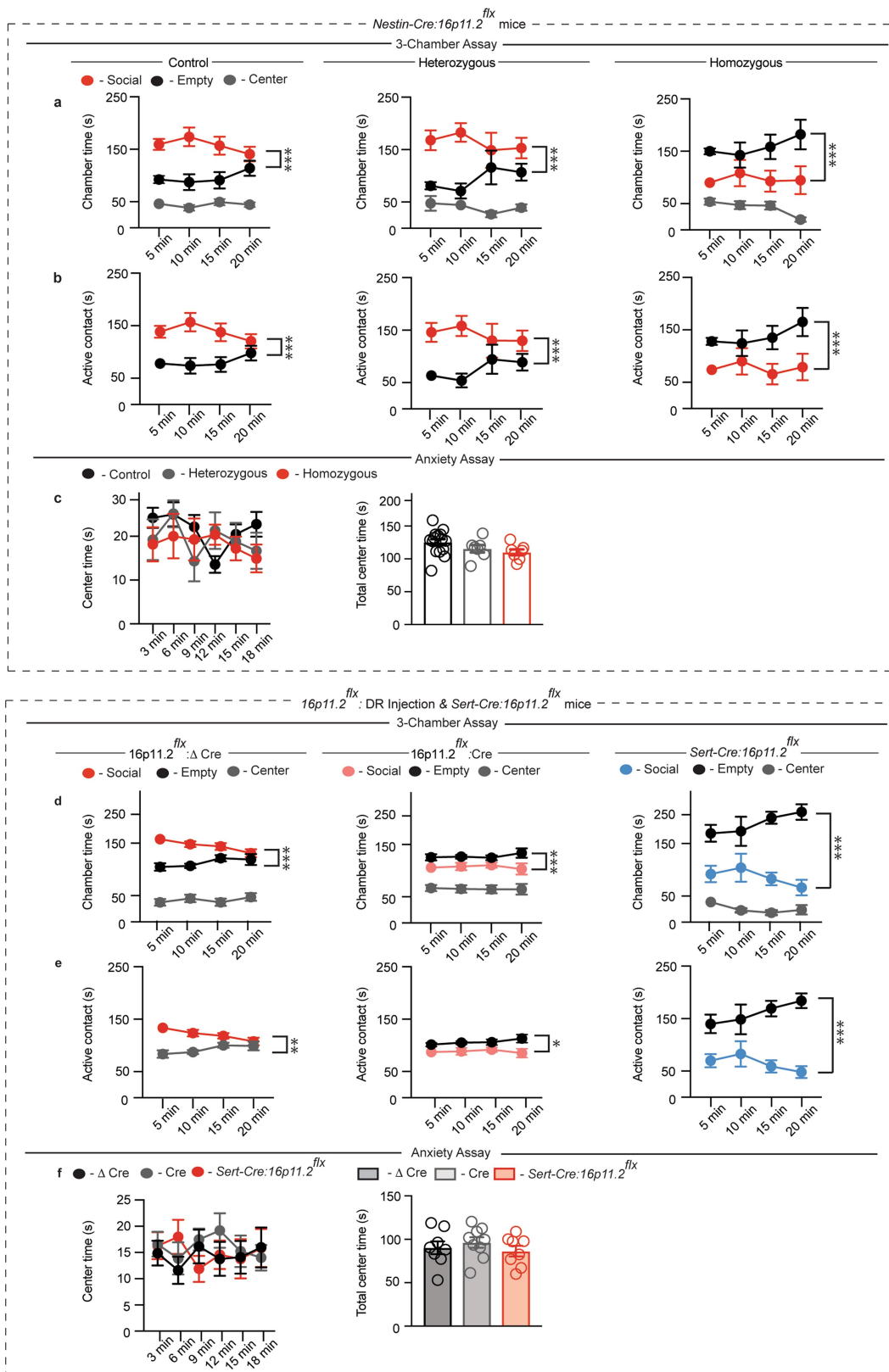
chamber assay. **i, j**, Quantification of chamber time (**i**: eYFP, $F_{6,54} = 1.058$, $P = 0.3989$, $n = 7$; NpHR, $F_{6,81} = 13.04$, $P < 0.01$, $n = 10$) and proximity time (**j**: eYFP, $F_{3,36} = 1.661$, $P = 0.1926$, $n = 7$; NpHR, $F_{3,54} = 16.29$, $P < 0.01$, $n = 10$) in the three-chamber assay. **k, l**, Quantification of novel object interaction assay (**k**: $F_{1,52} = 0.01018$, $P = 0.9200$, $n = 13-15$), locomotion assay (**l**: $F_{1,52} = 0.7626$, $P = 0.3865$, $n = 13-15$), and centre time (**m**: $F_{5,130} = 0.766$, $P = 0.5759$, $n = 13-15$) in *Sert-cre* mice expressing DIO-eYFP or DIO-ChR2 in DR receiving soma stimulation. **n-p**, Quantification of novel object interaction assay (**n**: $F_{1,30} = 0.04112$, $P = 0.8407$, $n = 7-10$), locomotion assay (**o**: $F_{1,30} = 0.3837$, $P = 0.5403$, $n = 7-10$), and centre time (**p**: $F_{5,80} = 1.195$, $P = 0.3190$, $n = 8-10$) in *Sert-cre* mice expressing DIO-eYFP or DIO-NpHR in DR receiving soma stimulation. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.



Extended Data Fig. 2 | Bidirectional modulation of DR-to-NAc 5-HT terminals modifies sociability, but not control behaviours.

a, b, Quantification of chamber time (**a**: eYFP, $F_{6,63} = 1.383$, $P = 0.2352$, $n = 8$; ChR2, $F_{6,72} = 4.891$, $P < 0.05$, $n = 9$) and proximity time (**b**: eYFP, $F_{3,42} = 0.9652$, $P = 0.4181$, $n = 8$; ChR2, $F_{3,48} = 7.565$, $P < 0.05$, $n = 9$) in the three-chamber assay. **c, d**, Quantification of chamber time (**c**: eYFP, $F_{6,81} = 1.626$, $P = 0.1506$, $n = 10$; NpHR, $F_{6,81} = 6.253$, $P < 0.05$, $n = 10$) and proximity time (**d**: eYFP, $F_{3,54} = 2.304$, $P = 0.0872$, $n = 10$; NpHR, $F_{3,54} = 7.821$, $P < 0.05$, $n = 10$) in the three-chamber assay. **e–g**, Quantification of novel object interaction assay (**e**: $F_{1,30} = 0.00206$, $P = 0.9641$, $n = 8–9$), locomotion assay (**f**: $F_{1,30} = 0.03023$, $P = 0.8631$,

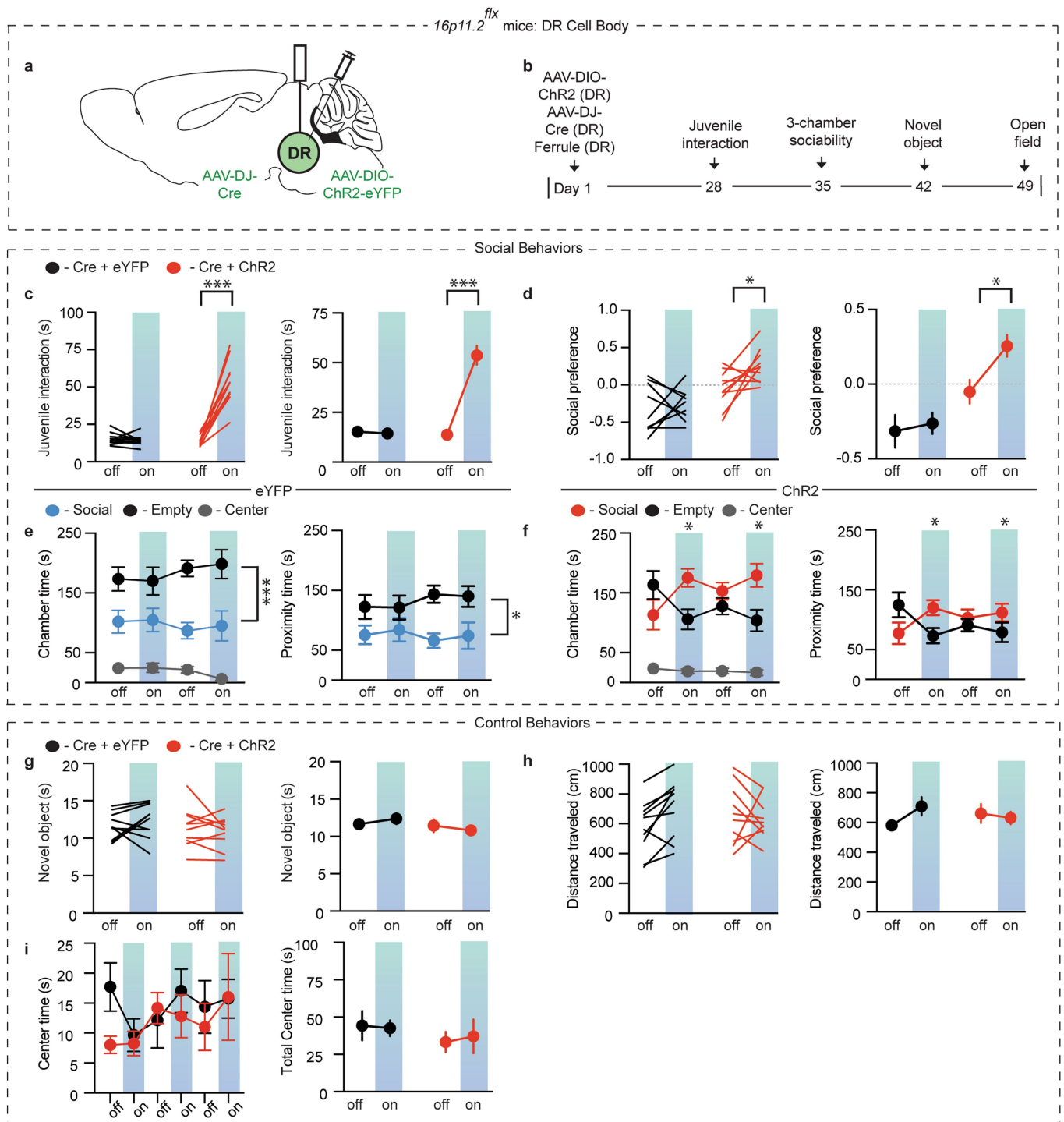
$n = 8–9$), and centre time (**g**: $F_{5,75} = 1.205$, $P = 0.3151$, $n = 8–9$) in Sert-cre mice expressing DIO-eYFP or DIO-ChR2 in DR receiving DR-to-NAc terminal stimulation. **h–j**, Quantification of novel object interaction assay (**h**: $F_{1,38} = 0.213$, $P = 0.6471$, $n = 10–11$), locomotion assay (**i**: $F_{1,34} = 1.077$, $P = 0.3066$, $n = 9–10$), and centre time (**j**: $F_{5,90} = 0.1646$, $P = 0.9749$, $n = 9–10$) in Sert-cre mice expressing DIO-eYFP or DIO-NpHR in DR receiving DR-to-NAc terminal stimulation. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | 16p11.2 deletion decreases sociability, but not an anxiety-related behaviour. a, b, Quantification of chamber time (a: control, $F_{2,45} = 39.5$, $P < 0.001$, $n = 16$; heterozygous, $F_{2,21} = 23.39$, $P < 0.001$, $n = 8$; homozygous, $F_{2,21} = 31.54$, $P < 0.001$, $n = 8$) and proximity time (b: control, $F_{1,30} = 14.61$, $P < 0.001$, $n = 16$; heterozygous, $F_{1,14} = 10.14$, $P < 0.01$, $n = 8$; homozygous, $F_{1,14} = 11.88$, $P < 0.01$, $n = 8$) in the three-chamber assay. **c,** Quantification of centre time (c: $F_{10,135} = 1.03$, $P = 0.4215$, $n = 7-15$) in control, heterozygous $16p11.2^{flx};Nes-creER$ and homozygous $16p11.2^{flx};Nes-creER$ mice. **d, e,** Quantification of chamber time in the three-chamber assay (d: $16p11.2^{flx};\Delta cre$, $F_{2,24} = 121.2$,

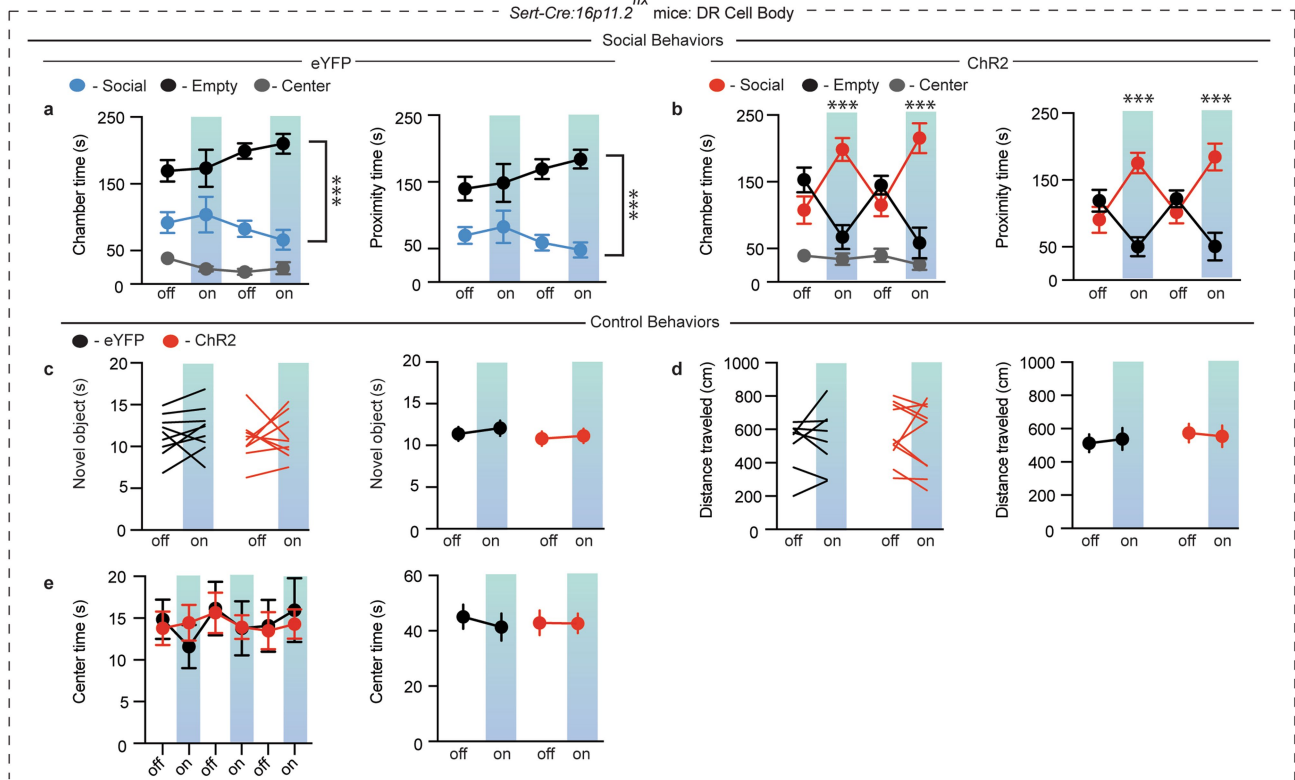
$P < 0.001$, $n = 9$; $16p11.2^{flx};cre$, $F_{2,24} = 27.86$, $P < 0.001$, $n = 9$; $Sert-cre:16p11.2^{flx}$, $F_{2,21} = 84.95$, $P < 0.001$, $n = 8$) and proximity time (e: $16p11.2^{flx};\Delta cre$, $F_{1,16} = 26.13$, $P < 0.001$, $n = 9$; $16p11.2^{flx};cre$, $F_{1,16} = 6.885$, $P < 0.05$, $n = 9$; $Sert-cre:16p11.2^{flx}$, $F_{1,14} = 44.00$, $P < 0.001$, $n = 8$). **f,** Quantification of centre time in (c: $F_{2,22} = 0.6019$, $P = 0.5565$, $n = 8-9$) $16p11.2^{flx};\Delta cre$, $16p11.2^{flx};cre$ and $Sert-cre:16p11.2^{flx}$ mice. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.



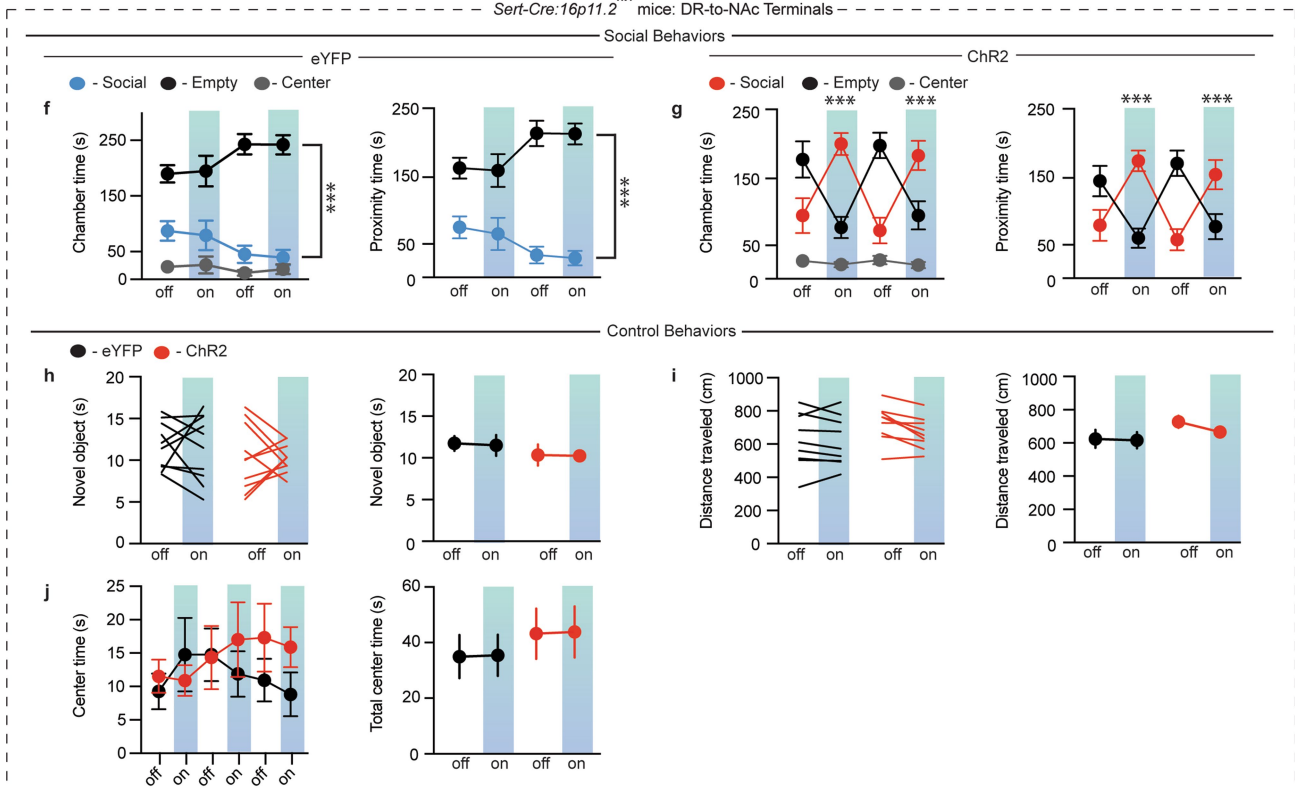
Extended Data Fig. 4 | Optogenetic activation of DR neurons reverses social deficits induced by 16p11.2 deletion, but does not alter control behaviours. **a**, Schematic of AAV-DJ-Cre and DIO-ChR2 injected into and optic fibre implanted above the DR in 16p11.2^{flx} mice. **b**, Timeline of behavioural experiments. **c, d**, Quantification of sociability during juvenile interaction (c: $F_{1,36} = 62.43$, $P < 0.001$, $n = 10$) and the three-chamber sociability assay (d: $F_{1,34} = 2.298$, $P < 0.05$, $n = 9-10$) in 16p11.2^{flx} mice expressing DIO-eYFP or DIO-ChR2 and AAV-DJ-Cre in DR receiving soma stimulation. **e, f**, Quantification of chamber and proximity time in the three-chamber assay (e: eYFP, $F_{2,24} = 92.48$, $P < 0.001$, $n = 9$ (left);

$F_{1,16} = 17.53$, $P < 0.05$, $n = 9$ (right); f: ChR2, $F_{6,81} = 3.085$, $P < 0.05$, $n = 10$ (left); $F_{3,54} = 3.493$, $P < 0.05$, $n = 10$ (right)). **g-i**, Quantification of novel object interaction assay (g: $F_{1,36} = 0.956$, $P = 0.3424$, $n = 10$), locomotion assay (h: $F_{1,36} = 1.962$, $P = 0.1698$, $n = 10$), and centre time (i: $F_{5,90} = 0.7668$, $P = 0.5761$, $n = 10$) in 16p11.2^{flx} mice. Data are mean ± s.e.m. * $P < 0.05$, *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

Sert-Cre:16p11.2^{flx} mice: DR Cell Body



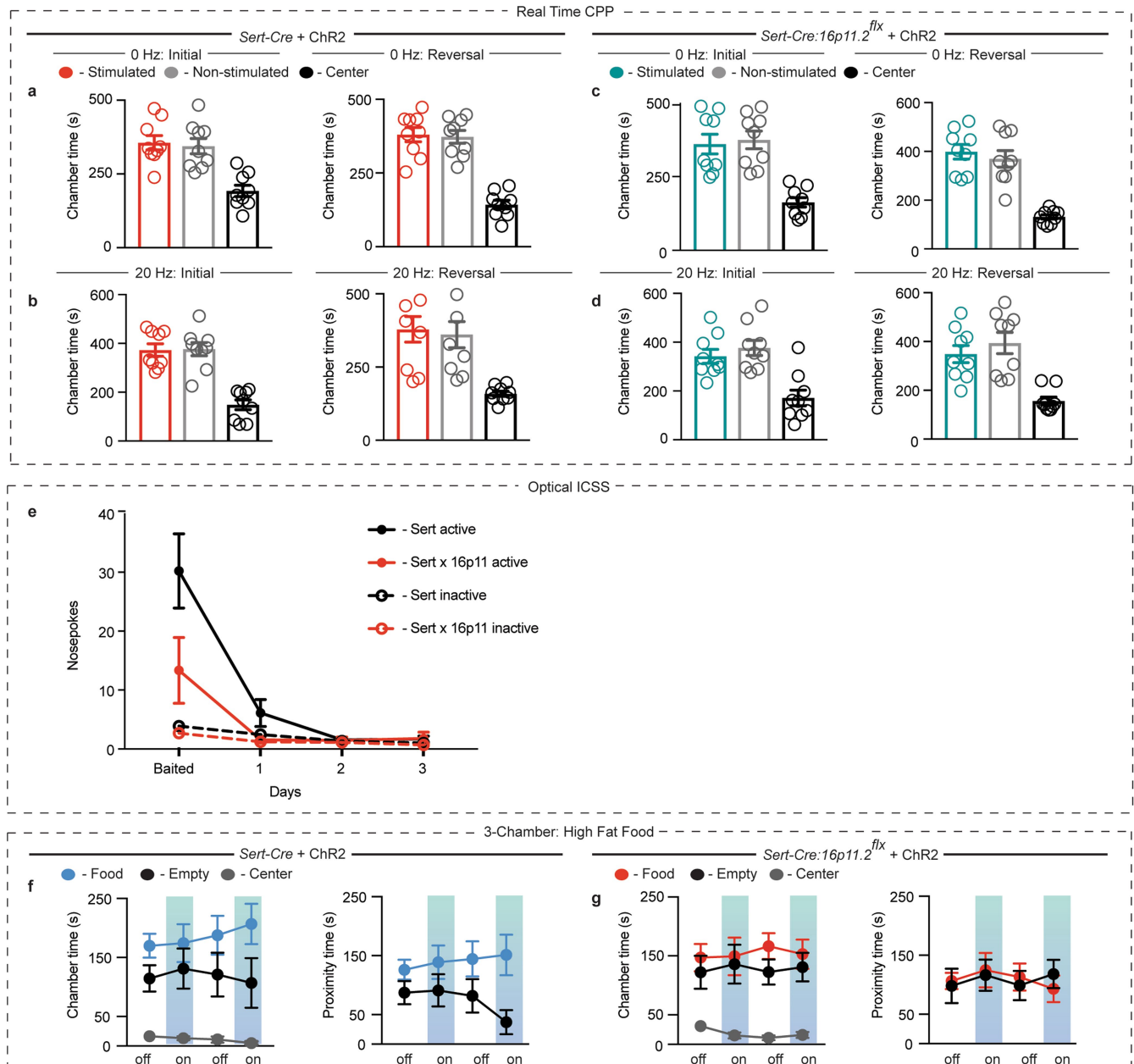
Sert-Cre:16p11.2^{flx} mice: DR-to-NAC Terminals



Extended Data Fig. 5 | See next page for caption.

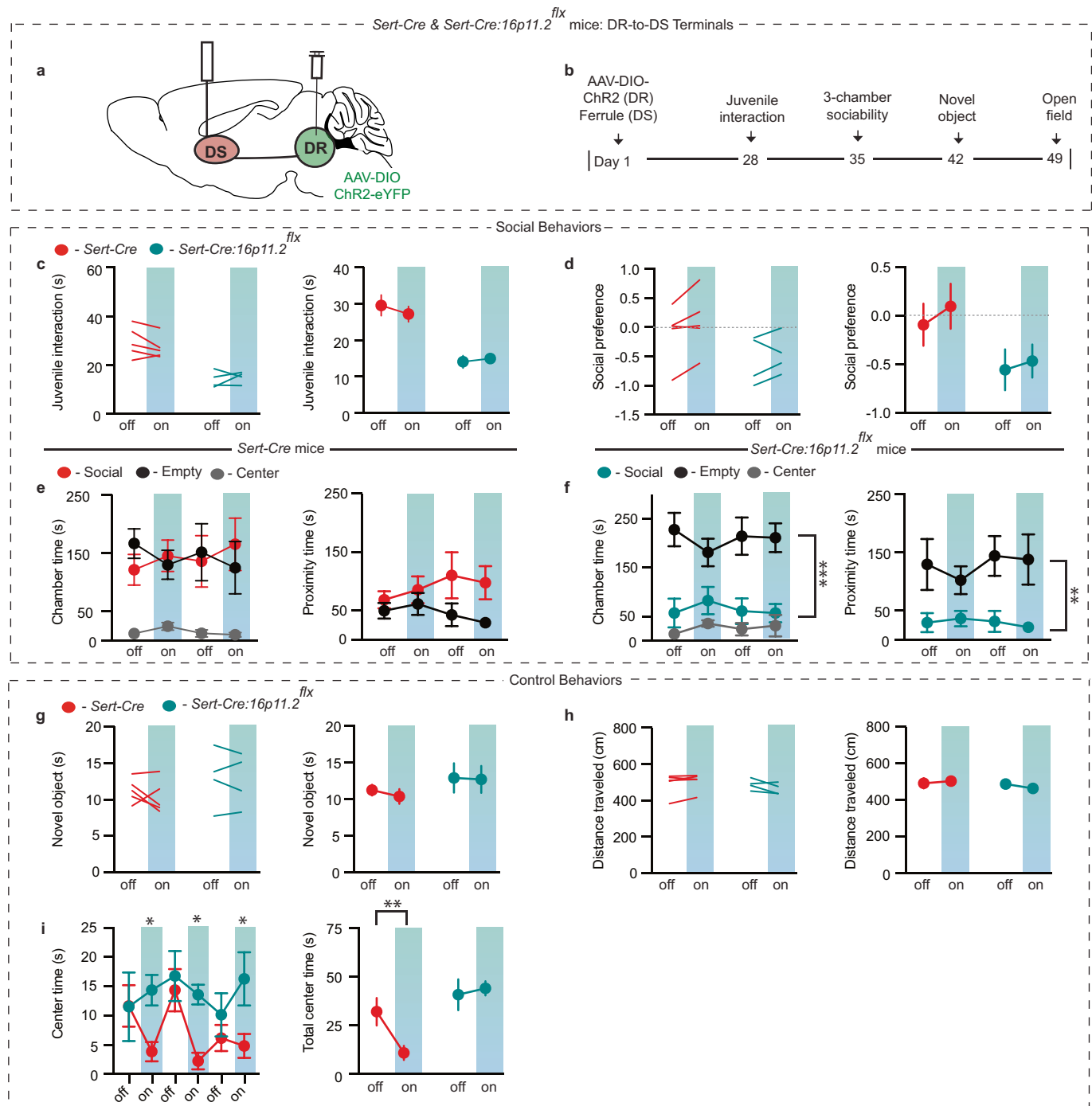
Extended Data Fig. 5 | Optogenetic activation of DR 5-HT neurons or DR-to-NAc 5-HT terminals rescues social deficits induced by 16p11.2 deletion, but does not alter control behaviours. **a, b**, Quantification of chamber and proximity time in the three-chamber assay (**a**: eYFP, $F_{2,21} = 84.95$, $P < 0.001$, $n = 8$ (left); $F_{1,14} = 44.00$, $P < 0.001$, $n = 8$ (right); **b**: ChR2, $F_{6,81} = 9.26$, $P < 0.001$, $n = 10$ (left); $F_{3,54} = 11.54$, $P < 0.001$, $n = 10$ (right)). **c–e**, Quantification of the novel object interaction assay (**c**: $F_{1,32} = 0.03819$, $P = 0.8463$, $n = 9$), locomotion assay (**d**: $F_{1,32} = 0.141$, $P = 0.7097$, $n = 8–10$), and centre time (**e**: $F_{5,80} = 0.195$, $P = 0.9636$, $n = 8–10$) in *Sert-cre:16p11.2^{flx}* mice expressing DIO-eYFP or DIO-ChR2 in DR receiving soma stimulation. **f, g**, Quantification of chamber

and proximity time in the three-chamber assay (**f**: eYFP, $F_{2,27} = 73.89$, $P < 0.001$, $n = 10$ (left); $F_{1,18} = 63.38$, $P < 0.001$, $n = 10$ (right); **g**: ChR2, $F_{6,81} = 11.33$, $P < 0.001$, $n = 10$ (left); $F_{3,54} = 14.55$, $P < 0.001$, $n = 10$ (right)). **h–j**, Quantification of novel object interaction assay (**h**: $F_{1,36} = 0.01038$, $P = 0.9194$, $n = 10$), locomotion assay (**i**: $F_{1,32} = 0.3655$, $P = 0.5497$, $n = 9$), and centre time (**j**: $F_{5,90} = 0.9092$, $P = 0.4788$, $n = 10$) in *Sert-cre:16p11.2^{flx}* mice expressing DIO-eYFP or DIO-ChR2 in DR receiving DR-to-NAc 5-HT terminal stimulation. Data are mean \pm s.e.m. *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.



Extended Data Fig. 6 | 5-HT release in the NAc is not acutely reinforcing. **a, b**, Quantification of chamber time in the real-time CPP assay for 0 Hz (**a**) and 20 Hz (**b**) stimulation for initial (left) and reversal (right) stimulations (**a**: 0 Hz initial $F_{2,16} = 10.22$, $P < 0.05$, $n = 9$; 0 Hz reversal $F_{2,16} = 29.01$, $P < 0.001$, $n = 9$; **b**: 20 Hz initial $F_{2,16} = 19.37$, $P < 0.001$, $n = 9$; 20 Hz reversal $F_{2,16} = 7.53$, $P < 0.01$, $n = 9$) in Sert-cre mice receiving DR-to-NAc terminal stimulation. **c, d**, Quantification of chamber time in the real-time CPP assay for 0 Hz (**c**) and 20 Hz (**d**) stimulation for initial (left) and reversal (right) stimulations (**c**: 0 Hz initial $F_{2,16} = 12.44$, $P < 0.001$, $n = 9$; 0 Hz reversal $F_{2,16} = 19.92$, $P < 0.001$, $n = 9$; **d**: 20 Hz initial $F_{2,16} = 8.56$, $P < 0.01$, $n = 9$; 20 Hz reversal $F_{2,16} = 9.517$, $P < 0.001$, $n = 9$) in Sert-cre:16p11.2^{flx} mice receiving DR-to-NAc terminal stimulation. **e**, Quantification of nose-pokes for active and inactive

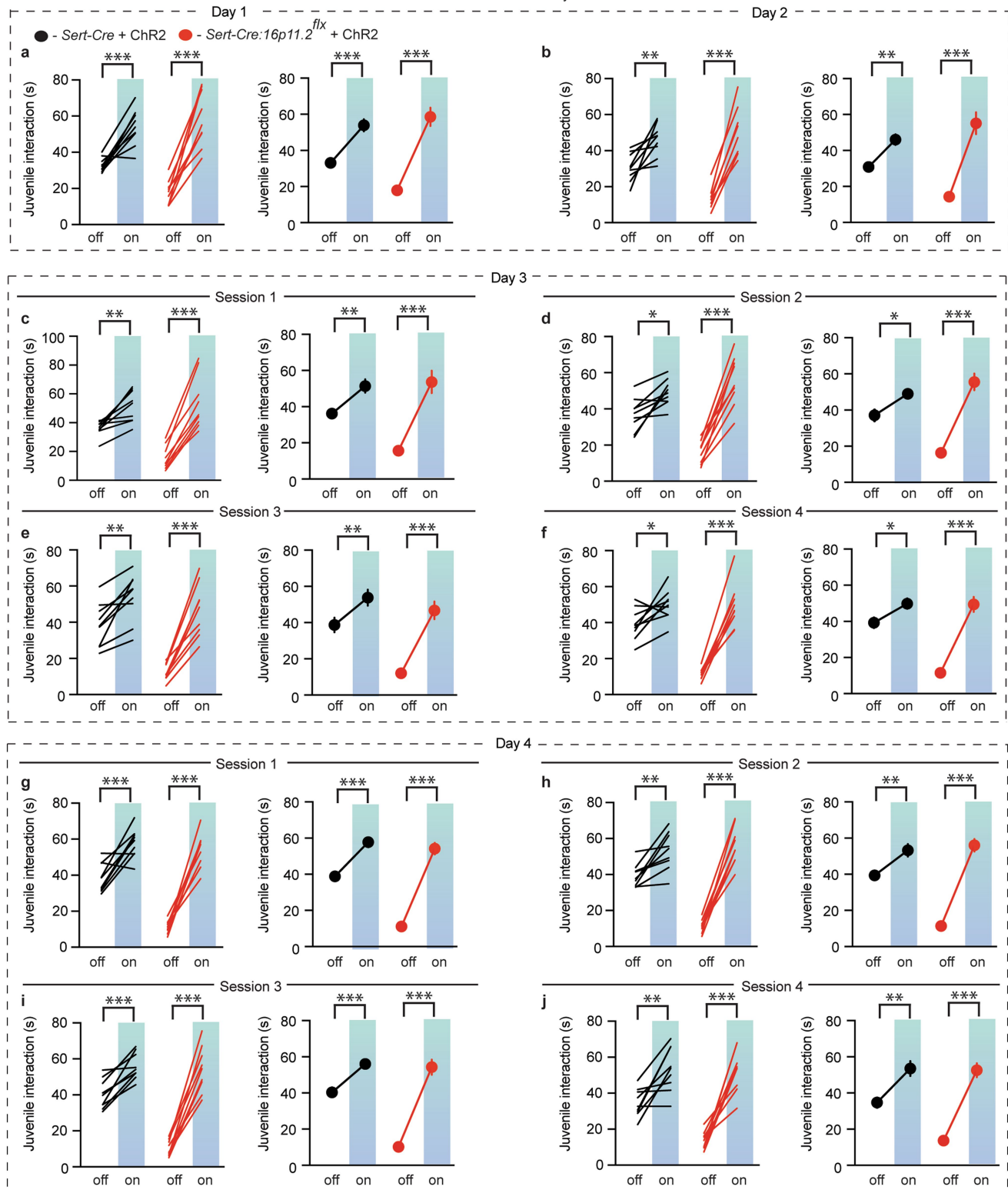
ports for Sert-cre and Sert-cre:16p11.2^{flx} mice (Sert-cre $F_{2,32} = 1.9$, $P = 0.1655$, $n = 9$; Sert-cre:16p11.2^{flx} $F_{2,32} = 0.25$, $P = 0.7821$, $n = 9$). **f, g**, Quantification of chamber time (left) and proximity time (right) in high-fat food three-chamber assay for Sert-cre mice ($F_{6,72} = 0.6713$, $P = 0.6731$, $n = 9$; $F_{3,48} = 1.495$, $P = 0.2279$, $n = 10$) and Sert-cre:16p11.2^{flx} mice ($F_{6,72} = 0.4006$, $P = 0.8763$, $n = 10$; $F_{3,48} = 0.4157$, $P = 0.7425$, $n = 10$) (**g**). Post-hoc analysis showed no significant difference between stimulated and non-stimulated chambers. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; repeated measures, one-way (**a–d**) or two-way (**f, g**) ANOVA with Tukey's multiple comparison post hoc test, or one-way ANOVA with Sidak's multiple comparison post hoc test comparing active to inactive nose-pokes (**e**). Comparisons with no asterisk had $P > 0.05$ and were considered not significant.



Extended Data Fig. 7 | Activation of DR 5-HT terminals in the dorsal striatum does not enhance sociability nor rescue social deficits induced by 16p11.2 deletion. **a**, Schematic of DIO-ChR2-eYFP injected into the DR and optic fibre implanted above the dorsal striatum (DS) in *Sert-cre* or *Sert-cre*:16p11.2^{flx} mice. **b**, Timeline of behavioural experiments. **c**, **d**, Quantification of sociability during the juvenile interaction assay (**c**: $F_{1,14} = 0.5429$, $P = 0.4734$, $n = 4-5$) and sociability in the three-chamber assay (**d**: $F_{1,14} = 0.055$, $P = 0.8177$, $n = 4-5$) in *Sert-cre* or *Sert-cre*:16p11.2^{flx} mice expressing DIO-eYFP or DIO-ChR2 with terminal stimulation of DR 5-HT neurons in the dorsal striatum. **e**, **f**, Quantification of the three-chamber assay showing chamber (left) and proximity (right)

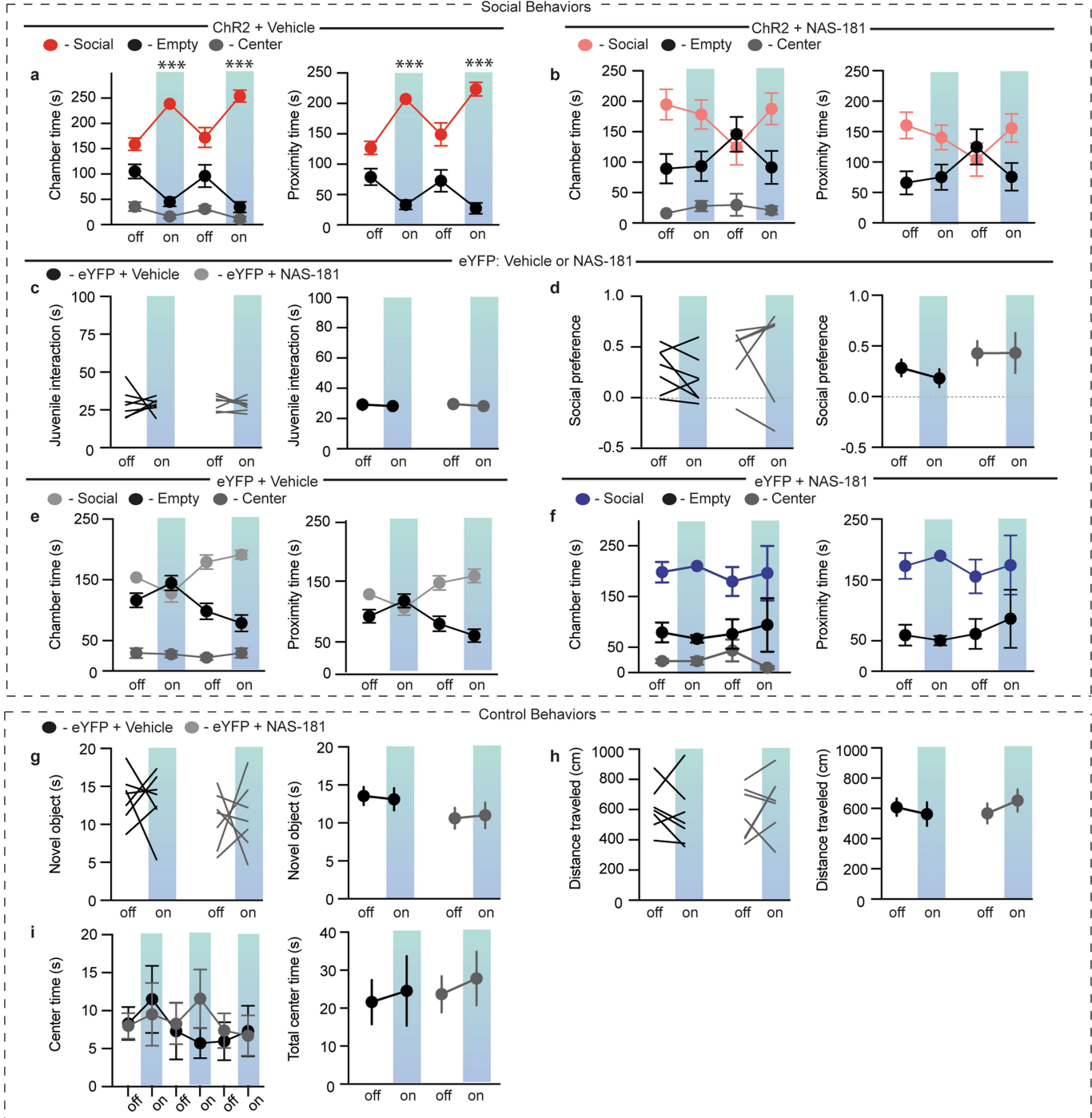
time (**e**: *Sert-cre*, $F_{6,36} = 1.078$, $P = 0.3939$, $n = 5$ (left); $F_{3,24} = 2.009$, $P = 0.1395$, $n = 5$ (right); **f**: *Sert-cre*:16p11.2^{flx}, $F_{2,9} = 16.44$, $P < 0.001$, $n = 4$ (left); $F_{1,6} = 11.2$, $P < 0.01$, $n = 4$ (right)). **g**–**i**, Quantification of novel object interaction assay (**g**: $F_{1,14} = 0.058$, $P = 0.8125$, $n = 4-5$), locomotion assay (**h**: $F_{1,14} = 0.6195$, $P = 0.4444$, $n = 4-5$), and centre time (**i**: $F_{1,7} = 8.292$, $P < 0.05$, $n = 4-5$) in *Sert-cre* or *Sert-cre*:16p11.2^{flx} mice. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and considered were not significant. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

Juvenile Interaction Assay



Extended Data Fig. 8 | Activation of DR-to-NAc 5-HT terminals does not elicit long-lasting effects on sociability. a–j, Quantification of sociability during juvenile interaction assay on day 1 of stimulation (a: $F_{1,16} = 10.65$, $P < 0.001$, $n = 9$), day 2 (b: $F_{1,16} = 16.71$, $P < 0.01$, $n = 9$), day 3 session 1 (c: $F_{1,16} = 17.7$, $P < 0.01$, $n = 9$), session 2 (d: $F_{1,16} = 27.05$, $P < 0.05$, $n = 9$), session 3 (e: $F_{1,16} = 13.35$, $P < 0.01$, $n = 9$), session 4 (f: $F_{1,16} = 25.03$, $P < 0.05$, $n = 9$) and day 4 session 1 (g: $F_{1,16} = 17.41$,

$P < 0.001$, $n = 9$), session 2 (h: $F_{1,16} = 39.76$, $P < 0.01$, $n = 9$), session 3 (i: $F_{1,16} = 35.49$, $P < 0.001$, $n = 9$), and session 4 (j: $F_{1,16} = 8.295$, $P < 0.01$, $n = 9$) in Sert-cre or Sert-cre:16p11.2^{flx} mice expressing DIO-ChR2 with DR-to-NAc 5-HT terminal stimulation. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test.

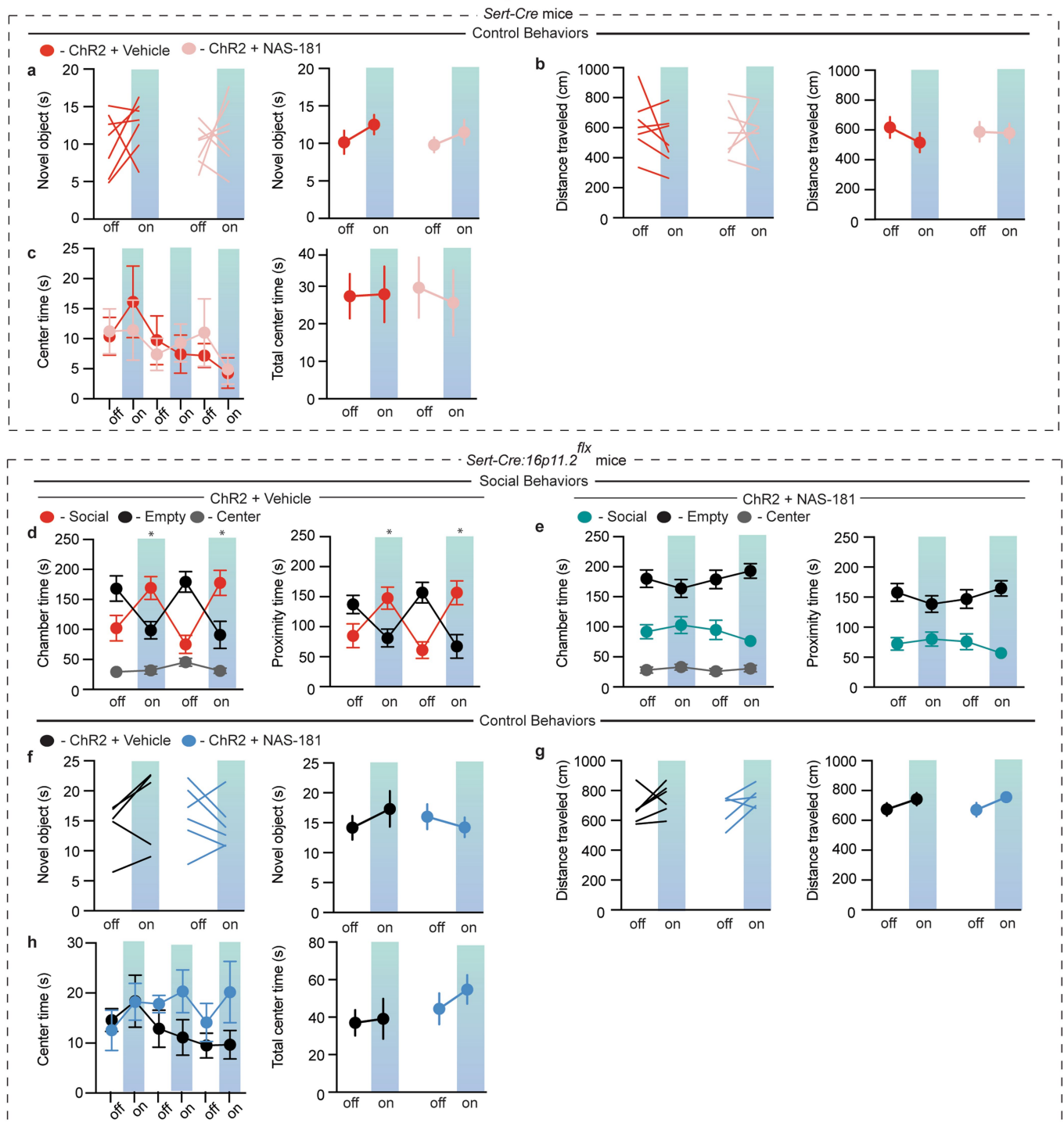


Extended Data Fig. 9 | 5-HT_{1b} receptor antagonist infusion into NAc blocks enhanced sociability due to DR 5-HT neuron stimulation.

a, Quantification of chamber and proximity time in the three-chamber assay (**a**: ChR2 + vehicle, $F_{6,54} = 15.15$, $P < 0.001$, $n = 7$ (left); $F_{3,36} = 18.00$, $P < 0.001$, $n = 7$ (right); **b**: ChR2 + NAS-181, $F_{6,45} = 1.479$, $P = 0.2069$, $n = 6$ (left); $F_{3,30} = 1.926$, $P = 0.1467$, $n = 6$ (right)) in *Sert-cre* mice expressing DIO-ChR2 in DR with either vehicle (red) or NAS-181 (pink) infused into the NAc before analysis of behaviour. **c**, Quantification of sociability during the juvenile interaction assay (**c**: $F_{1,24} = 0.004638$, $P = 0.9463$, $n = 7$) and the three-chamber assay (**d**: $F_{1,22} = 0.1686$, $P = 0.1686$, $n = 6-7$) in *Sert-cre* mice expressing DIO-eYFP in DR with either vehicle (black) or NAS-181 (grey) infused into NAc before behaviour. **e**, **f**, Quantification of the three-chamber assay showing

chamber and proximity time (**e**: eYFP + vehicle, $F_{6,54} = 12.36$, $P < 0.001$, $n = 7$ (left); $F_{3,36} = 15.98$, $P < 0.01$, $n = 7$ (right); **f**: eYFP + NAS-181, $F_{6,45} = 0.4512$, $P = 0.8403$, $n = 6$ (left); $F_{3,30} = 0.7365$, $P = 0.7365$, $n = 6$ (right)) in *Sert-cre* mice expressing DIO-eYFP in DR with either vehicle (grey) or NAS-181 (blue) infused into the NAc before analysis of behaviour. **g-i**, Quantification of the novel object interaction assay (**g**: $F_{1,24} = 0.0791$, $P = 0.7809$, $n = 7$), locomotion assay (**h**: $F_{1,24} = 0.8954$, $P = 0.3535$, $n = 7$), and centre time (**i**: $F_{5,60} = 0.6042$, $P = 0.6969$, $n = 7$) in *Sert-cre* mice expressing DIO-eYFP in DR with either vehicle (grey) or NAS-181 (blue) infused into the NAc before behaviour. Data are mean \pm s.e.m. *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.

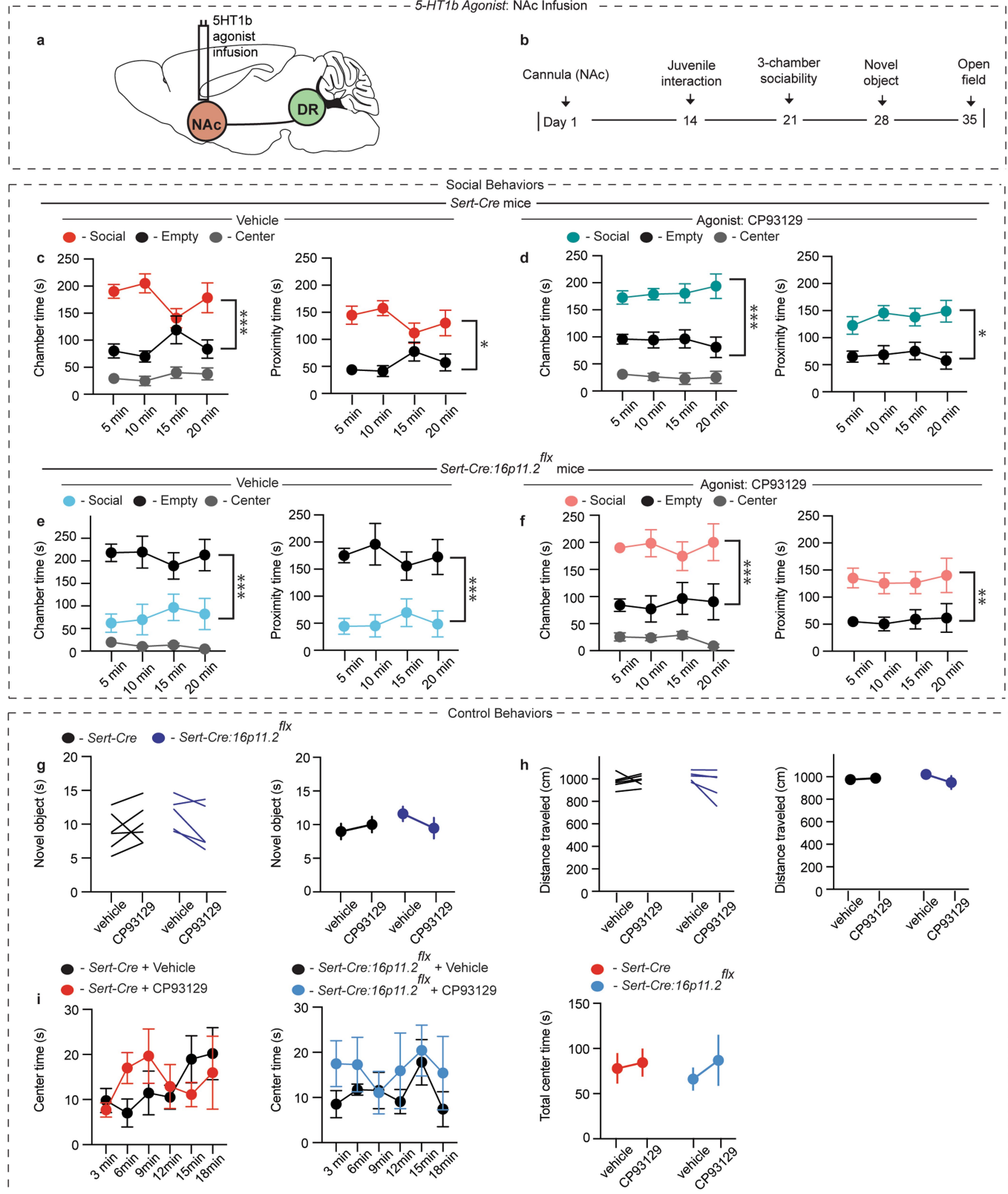
5-HT1b Antagonist: NAc Infusion



Extended Data Fig. 10 | 5-HT1b receptor antagonist infusion into the NAc does not alter control behaviours and blocks rescue of 16p11.2 deletion social deficits by DR 5-HT neuron stimulation.

a–c, Quantification of the novel object interaction assay (**a**: $F_{1,24} = 0.0579$, $P = 0.8120$, $n = 7$), locomotion assay (**b**: $F_{1,24} = 0.4764$, $P = 0.4967$, $n = 7$), and centre time (**c**: $F_{5,60} = 0.3936$, $P = 0.8513$, $n = 7$) in *Sert-cre* mice expressing DIO-ChR2 in DR with either vehicle (red) or NAS-181 (pink) infused into the NAc before behaviour. **d, e**, Quantification of chamber and proximity time in the three-chamber assay (**d**: ChR2 + vehicle, $F_{6,90} = 9.03$, $P < 0.05$, $n = 11$ (left); $F_{3,60} = 13.04$, $P < 0.05$, $n = 11$ (right); **e**: ChR2 + NAS-181, $F_{6,90} = 1.226$, $P = 0.3004$, $n = 11$ (left); $F_{3,60} = 1.69$,

$P = 0.1786$, $n = 11$ (right)) in *Sert-cre:16p11.2^{flx}* mice expressing DIO-ChR2 in DR with either vehicle (red) or NAS-181 (aqua) infused into the NAc before behaviour assays during optogenetic stimulation. **f–h**, Quantification of novel object interaction assay (**f**: $F_{1,18} = 0.263$, $P = 0.2758$, $n = 5–6$), locomotion assay (**g**: $F_{1,18} = 0.0344$, $P = 0.8549$, $n = 5–6$), and centre time (**h**: $F_{5,45} = 1.22$, $P = 0.3155$, $n = 5–6$) in *Sert-cre:16p11.2^{flx}* mice expressing DIO-ChR2 in DR with either vehicle (black) or NAS-181 (blue) infused into the NAc before behaviour assays. Data are mean \pm s.e.m. * $P < 0.05$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant.

5-HT_{1b} Agonist: NAc Infusion

Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Infusion of 5-HT1b receptor agonist into the NAc increases sociability, but does not alter control behaviours.

a, Schematic of 5-HT1b receptor agonist (CP93129) infusion into the NAc in *Sert-cre* or *Sert-cre:16p11.2^{flx}* mice. **b**, Timeline of behavioural experiments. **c, d**, Quantification of chamber and proximity time in the three-chamber assay (**c**: vehicle, $F_{2,15} = 45.82$, $P < 0.001$ (left); $F_{3,30} = 4.453$, $P < 0.05$ (right), $n = 6$; **d**: CP93129, $F_{2,15} = 63.11$, $P < 0.001$ (left); $F_{1,10} = 17.23$, $P < 0.05$, (right), $n = 6$) in *Sert-cre* mice with either vehicle (red) or CP93129 (aqua) infused into the NAc before analysis of behaviour. **e, f**, Quantification of chamber and proximity time in the three-chamber assay (**e**: vehicle, $F_{2,12} = 67.49$, $P < 0.001$ (left); $F_{1,8} = 41.03$, $P < 0.001$ (right), $n = 5$; **f**: CP93129, $F_{2,12} = 29.73$, $P < 0.001$ (left); $F_{1,8} = 17.17$,

$P < 0.01$ (right), $n = 5$) in *Sert-cre:16p11.2^{flx}* mice with either vehicle (blue) or CP93129 (pink) infused into the NAc before analysis of behaviour. **g–i**, Quantification of the novel object interaction assay (**g**: $F_{1,18} = 1.612$, $P = 0.2204$, $n = 5–6$), locomotion assay (**h**: $F_{1,18} = 0.0149$, $P = 0.9041$, $n = 5–6$), and centre time (**i**: $F_{5,50} = 1.492$, $P = 0.2093$; $F_{5,40} = 0.4561$, $P = 0.8063$, $n = 5–6$) in *Sert-cre* and *Sert-cre:16p11.2^{flx}* mice with either vehicle or NAS-181 infused into the NAc before behaviour. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; repeated measures, two-way ANOVA with Tukey's multiple comparison post hoc test. Comparisons with no asterisk had $P > 0.05$ and were considered not significant. The schematic of the mouse brain in this figure has been adapted with permission from Franklin & Paxinos⁴⁶.

Modulating plant growth–metabolism coordination for sustainable agriculture

Shan Li^{1,2}, Yonghang Tian¹, Kun Wu¹, Yafeng Ye¹, Jianping Yu¹, Jianqing Zhang¹, Qian Liu¹, Mengyun Hu³, Hui Li³, Yiping Tong¹, Nicholas P. Harberd⁴ & Xiangdong Fu^{1,2*}

Enhancing global food security by increasing the productivity of green revolution varieties of cereals risks increasing the collateral environmental damage produced by inorganic nitrogen fertilizers. Improvements in the efficiency of nitrogen use of crops are therefore essential; however, they require an in-depth understanding of the co-regulatory mechanisms that integrate growth, nitrogen assimilation and carbon fixation. Here we show that the balanced opposing activities and physical interactions of the rice GROWTH-REGULATING FACTOR 4 (GRF4) transcription factor and the growth inhibitor DELLA confer homeostatic co-regulation of growth and the metabolism of carbon and nitrogen. GRF4 promotes and integrates nitrogen assimilation, carbon fixation and growth, whereas DELLA inhibits these processes. As a consequence, the accumulation of DELLA that is characteristic of green revolution varieties confers not only yield-enhancing dwarfism, but also reduces the efficiency of nitrogen use. However, the nitrogen-use efficiency of green revolution varieties and grain yield are increased by tipping the GRF4–DELLA balance towards increased GRF4 abundance. Modulation of plant growth and metabolic co-regulation thus enables novel breeding strategies for future sustainable food security and a new green revolution.

The green revolution of the 1960s boosted crop yields, and was partly driven by widespread adoption of semi-dwarf green revolution varieties of cereals (GRVs)^{1–4}. GRV semi-dwarfism is due to the accumulation of growth-repressing DELLA proteins (DELLAs) conferred by mutant alleles at the *Rht* (wheat)^{5,6} and *SD1* (rice)^{7,8} loci. In normal plants, gibberellin (GA) promotes growth by stimulating the destruction of DELLAs^{9,10}. Mutant wheat GRV DELLAs⁵ are resistant to GA-stimulated destruction, whereas the rice GRV mutant *sd1* allele reduces bioactive GA abundance^{11,12}, thus increasing accumulation of the DELLA protein SLR1 (Fig. 1a, b). The conferred semi-dwarfism causes GRV resistance to yield-reducing ‘lodging’ (flattening of plants by wind and rain)⁴.

GRV lodging resistance is enhanced by relative insensitivity to nitrogen. For example, the nitrogen-induced increase in Nanjing6 (NJ6) plant height is reduced in NJ6-*sd1* (Fig. 1c), and the *Rht-B1b* GRV allele confers similar properties on wheat (Fig. 1d). Although DELLA accumulation inhibits GRV growth nitrogen response, nitrogen allocation to grain continues, thus combining enhanced harvestable yield with reduced lodging risk from increased nitrogen supply^{1,4,5,7,8}. These properties drove the rapid spread of GRV cultivation over the past 50 years³, and also ensured retention of semi-dwarfing alleles in current elite varieties^{5,6,12}. However, GRVs are associated with reduced nitrogen-use efficiency (NUE)¹³. Accordingly, mutant *sd1* and *Rht* alleles inhibit nitrogen uptake. For example, ammonium (NH₄⁺) is the majority nitrogen source for anaerobic paddy-field rice roots¹⁴. Although NJ6 ¹⁵NH₄⁺ uptake is regulated by nitrogen (the uptake rate is reduced by increasing nitrogen supply), *sd1* reduces the underlying NJ6-*sd1* uptake rate, and also interferes with its nitrogen-responsive regulation (Fig. 1e). Similarly, with nitrate (NO₃⁻) being the majority nitrogen source in aerobic soils¹⁵, the mutant *Rht-B1b* allele affects both underlying and nitrogen-regulated ¹⁵NO₃⁻ uptake in wheat (Fig. 1f). Thus, DELLA accumulation confers combined semi-dwarfism, reduced growth nitrogen response and reduced nitrogen uptake to GRVs. In consequence, achievement of high GRV yield requires environmentally damaging

nitrogen fertilizer inputs¹⁶. Development of new GRVs that combine high yields with reduced nitrogen supply is thus an urgent goal for global sustainable agriculture^{2,17}. We therefore analysed GRV growth–metabolism integration, reasoning that our discoveries might in turn enable development of new GRVs with improved NUE.

GRF4 promotes rice GRV ammonium uptake

We found approximately threefold variation in the ¹⁵NH₄⁺ uptake rates of 36 *sd1*-containing *indica* rice varieties and the *SD1*-containing NJ6 control (Fig. 2a), then crossed NM73 (having the highest rate; Fig. 2a) with NJ6 (recurrent parent) to generate a BC₁F₂ population. Quantitative trait locus (QTL) analysis of ¹⁵NH₄⁺ uptake rates revealed two logarithm of odds (LOD)-score peaks (quantitative trait loci *NGR1* and *NGR2* (*qNGR1* and *qNGR2*), Fig. 2b; Supplementary Table 1). Although the NM73 *qngr1* allele coincides in map position with *sd1*^{7,8}, the molecular identity of the NM73 *qngr2* allele, which was associated with increased ¹⁵NH₄⁺ uptake rates, was unknown. Positional mapping localized *qngr2* to *GRF4*^{18–20} (Extended Data Fig. 1a), suggesting a previously unknown function in NH₄⁺ uptake regulation. Because a NM73 (*GRF4*^{*nggr2*}) allele heterozygote has a higher rate than a NJ6 (*GRF4*^{*NGR2*}) allele homozygote (Extended Data Fig. 1b), *GRF4*^{*nggr2*} semi-dominantly increases NH₄⁺ uptakes. An NJ6-*GRF4*^{*nggr2*} isogenic line accordingly exhibited increased NH₄⁺ uptake rates (versus NJ6; Fig. 2c), and increased *GRF4* mRNA and GRF4 protein abundances (Fig. 2d, Extended Data Fig. 1c). Furthermore, RNA interference targeting *GRF4* reduced the high ¹⁵NH₄⁺ uptake rate of NJ6-*GRF4*^{*nggr2*}, whereas transgenic expression of *GRF4*^{*nggr2*} mRNA from its native promoter increased ¹⁵NH₄⁺ uptake (Fig. 2c, Extended Data Fig. 1c). Overexpression of either *GRF4*^{*NGR2*} or *GRF4*^{*nggr2*} mRNA from the constitutive rice *Actin1* promoter conferred increased ¹⁵NH₄⁺ uptake rates to NJ6 (Fig. 2c, Extended Data Fig. 1c). Thus, *GRF4*^{*nggr2*} is equivalent to *qngr2*, confers an increased ¹⁵NH₄⁺ uptake rate to NM73 and counteracts the repressive effects of *sd1*-mediated SLR1 accumulation.

¹State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. ²College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. ³Hebei Laboratory of Crop Genetics and Breeding, Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, China. ⁴Department of Plant Sciences, University of Oxford, Oxford, UK. *e-mail: xdfu@genetics.ac.cn

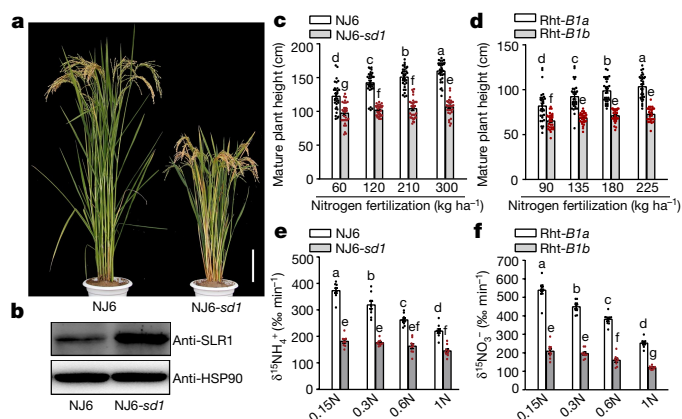


Fig. 1 | DELLA accumulation inhibits growth, nitrogen response and nitrogen uptake of rice and wheat GRVs. **a**, *Indica* rice variety NJ6 and near-isogenic NJ6-*sd1* plants. Scale bar, 15 cm. **b**, Accumulation of SLR1. Heat shock protein 90 (HSP90) serves as loading control. Blots are representative of three experiments performed independently with similar results. **c**, **d**, Heights of rice (**c**) and wheat (**d**) plants. Data are mean \pm s.e.m. ($n = 30$). **e**, $^{15}\text{NH}_4^+$ uptake rates in varying nitrogen supply (0.15N, 0.1875 mM NH_4NO_3 ; 0.3N, 0.375 mM NH_4NO_3 ; 0.6N, 0.75 mM NH_4NO_3 ; 1N, 1.25 mM NH_4NO_3). **f**, $^{15}\text{NO}_3^-$ uptake rates in varying nitrogen supply (0.15N, 0.1875 mM $\text{Ca}(\text{NO}_3)_2$; 0.3N, 0.375 mM $\text{Ca}(\text{NO}_3)_2$; 0.6N, 0.75 mM $\text{Ca}(\text{NO}_3)_2$; 1N, 1.25 mM $\text{Ca}(\text{NO}_3)_2$). Data in **e**, **f** are mean \pm s.e.m. ($n = 9$). **c**–**f**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test.

GRF4^{NGR2} (NJ6) and *GRF4*^{ngr2} (NM73) allelic comparisons revealed multiple single nucleotide polymorphisms (SNPs) (Extended Data Fig. 1a, d), two of which (g.1187T>A and g.1188C>A in exon 3)

prevent miR396-mediated cleavage of *GRF4*^{ngr2} mRNA^{18–20}, thus increasing *GRF4* mRNA and *GRF4* abundance (Fig. 2d, Extended Data Fig. 1c) and promoting $^{15}\text{NH}_4^+$ uptake. Nevertheless, variety RD23, which lacks 1187A and 1188A, also displays a high $^{15}\text{NH}_4^+$ uptake rate (Fig. 2a, Extended Data Fig. 1d), and shares three *GRF4* promoter SNPs (g.−884T>A, g.−847C>T and g.−801C>T; Extended Data Fig. 1a, d) with NM73. In all, we detected three *GRF4* promoter haplotypes (A, as in 9311 and other *indica* varieties; B, with −884A, −847T and −801T, as in NM73 and RD23; and C, common in *japonica* germplasm; Extended Data Fig. 1d). Notably, *GRF4* mRNA abundance is higher in haplotype B-containing varieties TZLL1 and RD23 (Extended Data Fig. 1d) than in elite varieties carrying haplotypes A or C (Extended Data Fig. 1e, f), and presumably confers their relatively high $^{15}\text{NH}_4^+$ uptake rates (Fig. 2a). Thus, NM73 has the highest of all assayed $^{15}\text{NH}_4^+$ uptake rates because it combines the effects of promoter haplotype B with the miR396 resistance conferred by 1187A and 1188A^{18–20}.

We also found that in addition to regulating $^{15}\text{NH}_4^+$ uptake, *GRF4* is regulated by nitrogen supply. NJ6 *GRF4* mRNA abundance decreases with increasing nitrogen (Fig. 2e), probably owing to decreased *GRF4* transcription (miR396 abundance is not detectably increased with increasing nitrogen; Extended Data Fig. 1g), thus reducing *GRF4* abundance (Fig. 2f). Because increased *GRF4* abundance increases $^{15}\text{NH}_4^+$ uptake (Fig. 2c, d), our observations suggest that promotion of *GRF4* abundance by low nitrogen enables feedback regulation of nitrogen homeostasis. In particular, the increased *GRF4* mRNA abundance response to low nitrogen is significantly amplified in varieties carrying haplotype B (for example, TZLL1 and RD23; Extended Data Fig. 1f). Finally, a CRISPR–Cas9²¹-generated semi-dwarf *grf4* mutant (Fig. 2g) lacks *GRF4* (Fig. 2h, Extended Data Fig. 1a), and exhibits reduced $^{15}\text{NH}_4^+$ influx (Fig. 2i), reduced nitrogen-responsive regulation of $^{15}\text{NH}_4^+$ uptake (Fig. 2i) and reduced nitrogen-dependent biomass

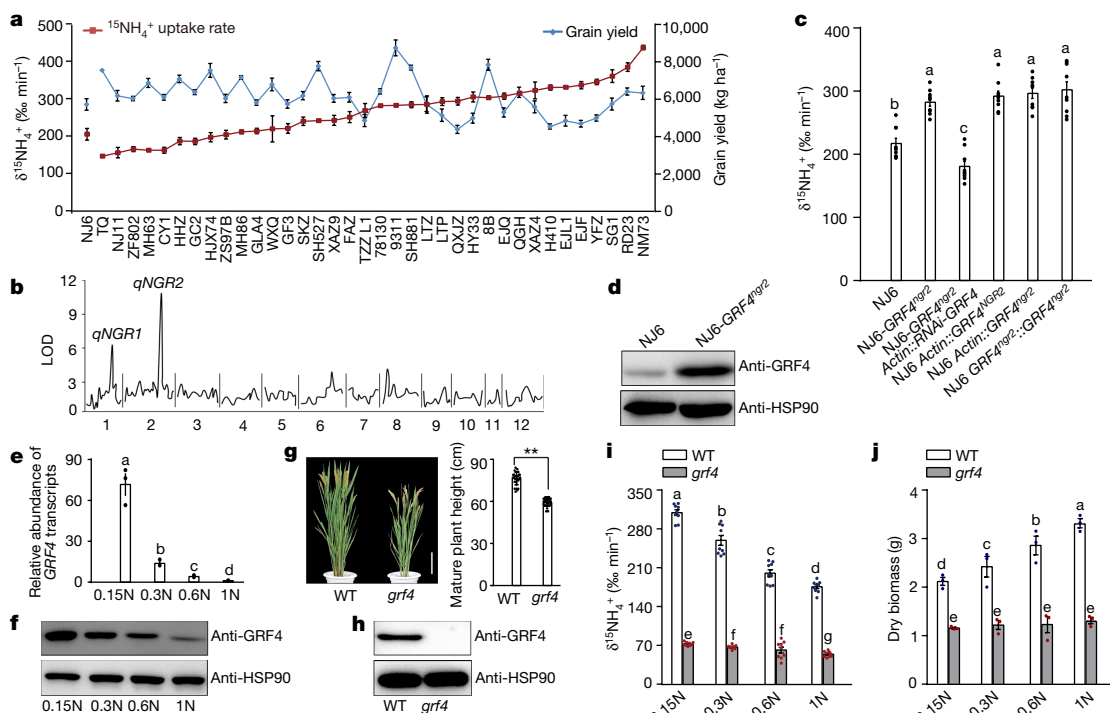


Fig. 2 | *GRF4* regulates rice NH_4^+ uptake and growth response to nitrogen availability. **a**, Variation in $^{15}\text{NH}_4^+$ uptake and grain yield. Four-week-old rice plants ($^{15}\text{NH}_4^+$ uptake assays) were grown hydroponically with high nitrogen supply (1.25 mM NH_4NO_3). Field-grown rice plants (yield assays) were grown with urea supply (210 kg ha^{-1}). Data are mean \pm s.e.m. of six plots (each plot contained 220 plants) per line. **b**, QTL analysis. **c**, $^{15}\text{NH}_4^+$ uptake rates. **d**, Accumulation of *GRF4*. **e**, *GRF4* transcript abundance in NJ6 roots grown in increasing nitrogen supply (0.15N, 0.1875 mM NH_4NO_3 ; 0.3N, 0.375 mM NH_4NO_3 ; 0.6N, 0.75 mM NH_4NO_3 ; 1N, 1.25 mM NH_4NO_3). Transcription is measured relative to

1N (set to one). **f**, Accumulation of *GRF4* in NJ6. **g**, Mature plant height of the rice *grf4* mutant (Extended Data Fig. 1a). Scale bar, 15 cm. Data are mean \pm s.e.m. ($n = 20$). $^{**}P < 0.05$ compared to the wild-type (WT) group using a two-sided Student's *t*-test. **h**, Accumulation of *GRF4*. HSP90 serves as loading control (**d**, **f**, **h**). **i**, $^{15}\text{NH}_4^+$ uptake rate. Data in **c**, **i** are mean \pm s.e.m. ($n = 9$). **j**, Dry weight of four-week-old plants. Data in **e**, **j** are mean \pm s.e.m. ($n = 3$). **c**, **i**, **j**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test. **b**, **d**, **f**, **h**, Data are representative of three experiments performed independently with similar results.

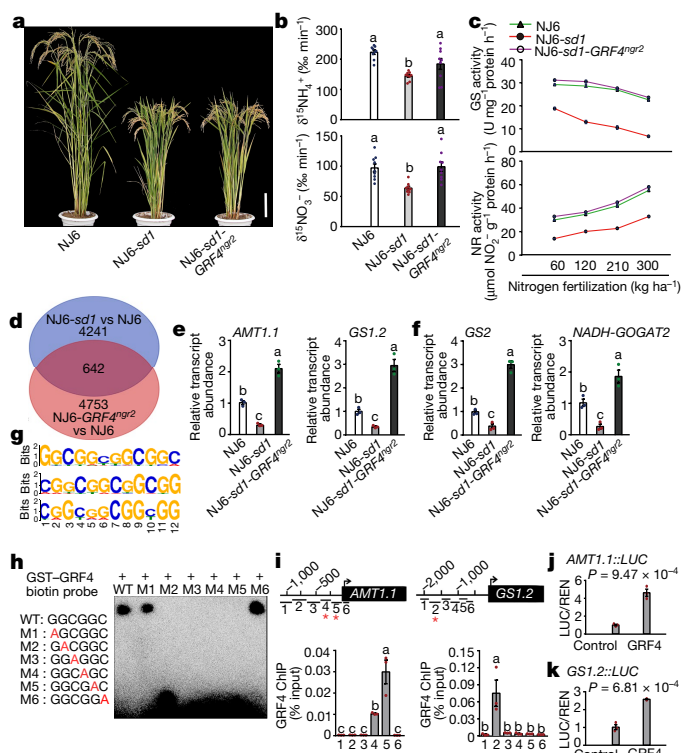


Fig. 3 | GRF4 regulates expression of multiple nitrogen-metabolism genes. **a**, Mature plants. Scale bar, 15 cm. **b**, $^{15}\text{NH}_4^+$ and $^{15}\text{NO}_3^-$ uptake rates. Data are mean \pm s.e.m. ($n = 9$). **c**, Glutamine synthase (GS) and nitrate reductase (NR) activities in shoots of rice plants grown in paddy-field conditions with increasing urea supply. **d**, RNA-seq analysis: 4,883 genes were downregulated in NJ6-*sd1* (versus NJ6; blue), 5,395 genes were upregulated in NJ6-GRF4^{ng2} (versus NJ6; orange) and 642 genes were common to both. **e**, **f**, Root (**e**) and shoot (**f**) mRNA abundances relative to NJ6 (set to one). **g**, Sequence motifs enriched in ChIP-seq with Flag-tagged GRF4. **h**, EMSA assays. Pictures in **a**, **h** are representative of three experiments performed independently with similar results. **i**, Flag-GRF4 mediated ChIP-PCR enrichment (relative to input) of GCGG-containing promoter fragments (marked with an asterisk) from *AMT1.1* and *GS1.2*. **b**, **e**, **f**, **i**, **j**, **k**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test. **j**, **k**, Transactivation assays. **c**, **e**, **f**, **i**, **k**, Data are mean \pm s.e.m. ($n = 3$). P values are from a two-sided Student's t -test.

accumulation (Fig. 2j). Thus, GRF4 is a nitrogen-responsive transcriptional regulator promoting both NH_4^+ uptake and growth in response to nitrogen supply, and counteracting the inhibitory effects of SLR1.

Regulation of nitrogen metabolism by GRF4–SLR1

We next determined how GRF4 and SLR1 counteract one another to regulate NH_4^+ assimilation. Although a NJ6-*sd1*-GRF4^{ng2} isogenic line retains the semi-dwarfism, tiller numbers per plant and grain numbers per panicle conferred by *sd1* (Fig. 3a, Extended Data Fig. 2a–c), leaf and culm width and grain yield are increased (Extended Data Fig. 2d–f). In addition, the $^{15}\text{NH}_4^+$ uptake rate in NJ6-*sd1*-GRF4^{ng2} is greater than in NJ6-*sd1* (and similar to that of NJ6) and $^{15}\text{NO}_3^-$ uptake is similarly affected (Fig. 3b). Furthermore, the activities of key nitrogen-assimilation enzymes, such as glutamine synthase (NH_4^+ assimilation)²² and nitrate reductase (NO_3^- assimilation)²³ are, at varying nitrogen supply levels, consistently greater in NJ6-*sd1*-GRF4^{ng2} than in NJ6-*sd1*, and similar to that of NJ6 (Fig. 3c). Thus, GRF4 promotes both nitrogen uptake and nitrogen assimilation, whereas SLR1 inhibits these processes.

Transcriptome-wide RNA sequencing (RNA-seq) analysis identified 642 genes with transcripts that were upregulated by GRF4 in NJ6-GRF4^{ng2} and downregulated by SLR1 in NJ6-*sd1* (versus NJ6) (Fig. 3d, Supplementary Tables 2, 3). Among these, quantitative PCR with reverse transcription (RT-qPCR) confirmed root abundances of mRNAs encoding NH_4^+ -uptake transporters (for example, *AMT1.1* and

AMT1.2)²⁴ to be increased in NJ6-*sd1*-GRF4^{ng2}, but reduced in NJ6-*sd1* (Fig. 3e, Extended Data Fig. 2g). Similarly, abundances of mRNAs encoding NH_4^+ -assimilation enzymes (for example, *GS1.2*²³, *GS2* and *NADH-GOGAT2*) and corresponding enzymatic activities were relatively enhanced in NJ6-*sd1*-GRF4^{ng2} (Fig. 3c, e, f, Extended Data Fig. 2h–j). Next, DNA sequencing of GRF4 chromatin-immunoprecipitation products (ChIP-seq) revealed potential GRF4 target-recognition sites, with a predominant GCGGC motif being common to multiple nitrogen-metabolism gene promoters (Fig. 3g, Supplementary Table 4). Electrophoretic mobility shift assays (EMSA) demonstrated binding of glutathione S-transferase-tagged GRF4 to DNA fragments containing intact but not mutant GCGG core motifs (Fig. 3h), and ChIP-PCR confirmed in vivo association of GRF4 with GCGG-containing promoter fragments from multiple NH_4^+ -metabolism genes, including *AMT1.1* and *GS1.2* (Fig. 3i, Extended Data Fig. 2k–n). Finally, GRF4 activates transcription from *AMT1.1* and *GS1.2* promoters in transactivation assays (Fig. 3j, k, Extended Data Fig. 2o). Further experiments demonstrated that GRF4-mediated transcriptional activation also promotes NO_3^- metabolism (Fig. 3b, c, Extended Data Fig. 3). Thus, GRF4 is a transcriptional activator of nitrogen metabolism, and counteracts the inhibitory effects of SLR1.

We next investigated how GA, SLR1 and GRF4 regulate nitrogen metabolism. GA promotes both $^{15}\text{NH}_4^+$ uptake rates to similarly high levels in NJ6 and NJ6-*sd1* (Fig. 4a). In addition, the GA-biosynthesis inhibitor paclobutrazol²⁵ (PAC) reduces $^{15}\text{NH}_4^+$ uptake in NJ6 and NJ6-*sd1*, whereas GA restores it (Fig. 4a). Thus, SLR1 accumulation (owing to *sd1* or PAC) reduces NH_4^+ uptake, whereas SLR1 reduction (owing to GA) increases it. Furthermore, the GA–DELLA system differentially regulates the abundance of NH_4^+ -metabolism mRNAs. *AMT1.1* and *GS1.2* mRNA abundances are increased by GA, reduced by PAC and restored by combined GA and PAC (Fig. 4b). We next found that PAC reduces ChIP-PCR enrichment of GCGG motif-containing fragments from the *AMT1.1* and *GS1.2* promoters, whereas GA promotes enrichment (Fig. 4c). Therefore, SLR1 accumulation inhibits binding of GRF4 to *AMT1.1* and *GS1.2* promoters (Fig. 4c), thus affecting mRNA abundance and NH_4^+ metabolism (Fig. 4a, b, Extended Data Fig. 4a, b), whereas SLR1 reduction promotes GRF4 binding. SLR1 abundance probably also affects NO_3^- uptake (Fig. 3b, Extended Data Fig. 3a, b) and nitrate reductase activity (Fig. 3c, Extended Data Fig. 4c) via inhibition of GRF4 activation of NO_3^- -metabolism genes.

Although the interaction of GRF4 with GIF (GRF-interacting factor) co-activators via a conserved QLQ domain (Extended Data Fig. 5a, b) promotes expression of target genes¹⁸, bimolecular fluorescence complementation (BiFC) and co-immunoprecipitation assays revealed that SLR1 interferes with this interaction (Fig. 4d–h, Extended Data Fig. 5c). In vivo fluorescence resonance energy transfer (FRET) assays demonstrated that SLR1 competitively inhibits the GRF4–GIF1 interaction, and that GA relieves this inhibition (Fig. 4f, g). Although the GRF4–GIF1 interaction promotes binding of GRF4 to GCGG motif-containing DNA fragments, SLR1 inhibits this promotion by inhibiting the GRF4–GIF1 interaction (but does not directly interfere with the DNA binding of GRF4; Fig. 4h). Accordingly, SLR1 inhibits GRF4–GIF1-mediated transactivation from *AMT1.1* and *GS1.2* promoters (Fig. 4i).

Notably, GRF4 abundance is self-promoted, and SLR1 inhibits that promotion. Although *GRF4* mRNA abundance is reduced in NJ6-*sd1* (versus NJ6) but increased in NJ6-*sd1*-GRF4^{ng2} (versus NJ6-*sd1*; Extended Data Fig. 6a), GA increases *GRF4* mRNA abundance, and overcomes PAC-mediated reductions in *GRF4* mRNA abundance (Extended Data Fig. 6b). Furthermore, GRF4 binds in vivo with GCGG-containing *GRF4* promoter fragments (Extended Data Fig. 6c), and SLR1 inhibits GRF4–GIF1-mediated transcriptional activation of the *GRF4* promoter (Extended Data Fig. 6d). In consequence, SLR1 reduces GRF4 abundance by interfering with the GRF4–GIF1 interaction, whereas the *GRF4*^{ng2} allele restores GRF4 abundance (Extended Data Fig. 6e; NJ6-*sd1* compared with NJ6-*sd1*-GRF4^{ng2}). Thus, interference of SLR1 with the GRF4–GIF1 interaction counteracts the promotive effects of GRF4 on nitrogen metabolism in two ways.

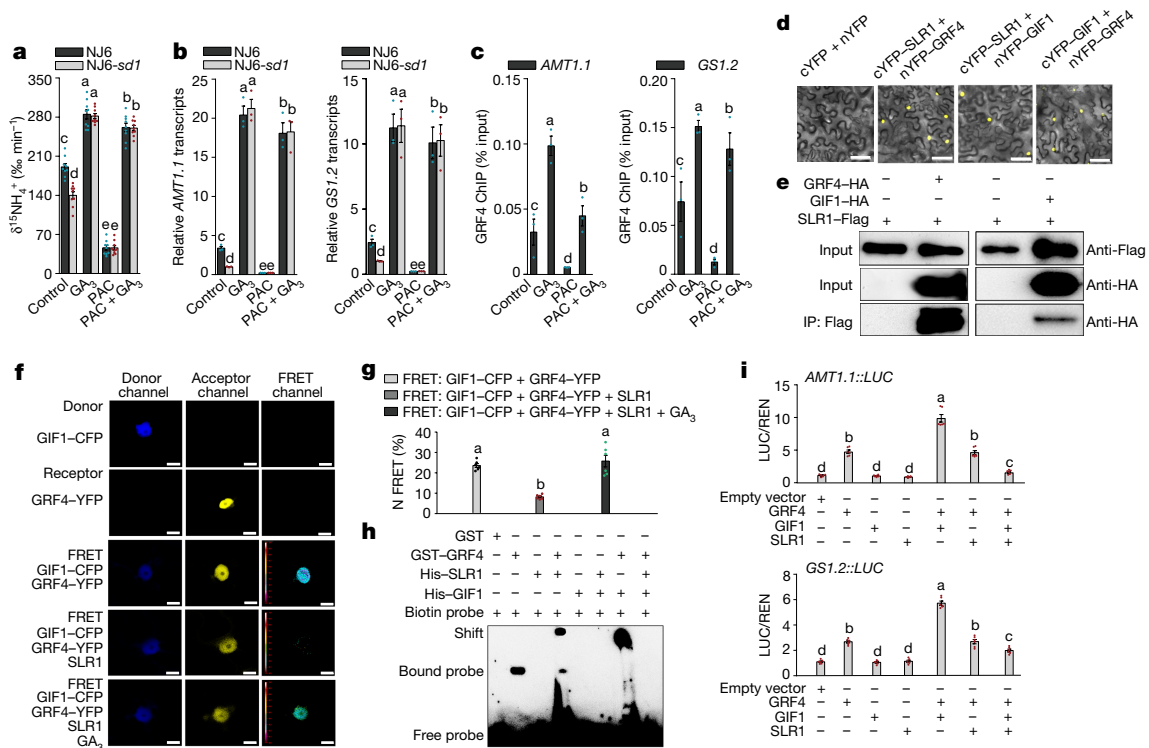


Fig. 4 | Competitive GRF4-GIF1-SLR1 interactions coordinate NH_4^+ uptake and assimilation. **a**, $^{15}\text{NH}_4^+$ uptake rates in four-week-old plants treated with 100 μM GA_3 and/or 2 μM paclobutrazol (PAC). Data are mean \pm s.e.m. ($n = 9$). **b**, Root mRNA abundance relative to the level in NJ6-*sd1* plants (set to one). **c**, Extent of ChIP-PCR GRF4-mediated enrichment (relative to input) of GCGG-containing promoter fragments from *AMT1.1* (fragment 5) and *GS1.2* (fragment 2) (shown in Fig. 3i). Data in **b**, **c** are mean \pm s.e.m. ($n = 3$). **d**, BiFC assays. Scale bar, 60 μm . cYFP, C-terminal portion of YFP; nYFP, N-terminal portion of YFP. When cYFP- and nYFP-tagged proteins are in close proximity, the cYFP and nYFP portions interact and produce YFP fluorescence. **e**, Co-

immunoprecipitation experiments. HA, haemagglutinin. **f**, FRET images. Scale bar, 200 μm . **g**, Mean normalized FRET (N FRET) data for GIF1-CFP and GRF4-YFP channels. **h**, EMSA assays. GST, glutathione S-transferase. **d-f**, **h**, Pictures are representative of three experiments performed independently with similar results. **i**, Transactivation assays. The luciferase (LUC)/renilla (REN) activity obtained from co-transfection with an empty effector construct and indicated reporter constructs. The activity of the empty effector construct was set to one. **g**, **h**, Data are mean \pm s.e.m. ($n = 6$). Different letters denote significant differences ($P < 0.05$) from Duncan's multiple range tests.

First, SLR1 reduces GRF4 accumulation. Second, SLR1 reduces GRF4-GIF1 activation of the transcription of nitrogen-metabolism genes.

GRF4-SLR1 links carbon fixation with growth

Although nitrogen metabolism is known to be coupled with the rate of photosynthetic carbon fixation²⁶, the molecular coupling mechanisms remain unknown. We next determined whether the GRF4-SLR1 interaction also regulates carbon assimilation. Transcriptome comparisons of NJ6, NJ6-*sd1* and NJ6-*sd1-GRF4^{gr2}* (Fig. 3d, Supplementary Tables 2, 3) indicated that GRF4 upregulates multiple genes encoding regulatory components of photosynthesis (for example, *Lhca1* and *CAB1*), sucrose metabolism (for example, *TPS1* and *TPP1*) and sucrose transport (for example, *SWEET11* and *SWEET12*) (Extended Data Fig. 7a, b), whereas SLR1 downregulates these same genes. In addition, GRF4 binds in vivo to GCGG-containing promoter fragments from *PsbS1*, *TPS1* and *SWEET11* (Extended Data Fig. 7c), whereas SLR1 inhibits GRF4-GIF1 activation of transcription from *PsbS1*, *TPS1* and *SWEET11* promoters (Extended Data Fig. 7d). For selected photosynthetic genes (*Lhca1*, *Lhca3*, *Lhca4*, *Lhcb2*, *PsaD* and *PsaE*), we confirmed that encoded protein abundances in NJ6, NJ6-*sd1* and NJ6-*sd1-GRF4^{gr2}* (Extended Data Fig. 7e) mirror the abundances of the respective encoding mRNAs (Extended Data Fig. 7a). Finally, we found that these effects on photosynthesis and carbon-assimilation gene expression affect carbon metabolism. First, in accordance with previous reports that semi-dwarfed GRVs have increased photosynthetic rates^{27,28}, we found the increased photosynthetic rate of NJ6-*sd1* (versus NJ6) to be still further increased in NJ6-*sd1-GRF4^{gr2}* (Extended Data Fig. 7f). Furthermore, reductions in NJ6 biomass and

carbon content conferred by *sd1* are reversed and further increased in NJ6-*sd1-GRF4^{gr2}* (Extended Data Fig. 7g, h), but without affecting the carbon:nitrogen (C:N) ratio (Extended Data Fig. 7i). Thus, antagonistic GRF4-SLR1 interaction regulates and coordinates both nitrogen and carbon assimilation (hence maintaining the C:N ratio).

We also found that GRF4 upregulates multiple genes promoting cell division, including those encoding cyclin-dependent *cdc2* protein kinases^{29,30} (for example, *cycA1;1* and *cdc20 s-3*; Extended Data Fig. 7j), whereas SLR1 again downregulates these genes. This finding is consistent with the plant growth reduction in the *grf4* mutant (Fig. 2g). In addition, GRF4 binds in vivo to GCGG-containing promoter fragments from *cycA1;1* and *cdc20 s-3* (Extended Data Fig. 7k), and GA promotes GRF4-GIF1 activation of transcription from these same promoters (Extended Data Fig. 7l), whereas SLR1 inhibits activation by GRF4-GIF1. We conclude that GRF4-SLR1 antagonism modulates the GA-mediated promotion of cell proliferation, and integrates growth, nitrogen and carbon metabolism regulation.

GRF4 increases GRV NUE and grain yield

GRF4 promoter haplotype B (Extended Data Fig. 1d) exists in selected *indica* cultivars, but not in modern elite varieties. Nevertheless, among 225 accessions³¹, haplotype B is associated with relatively high yield potential (Extended Data Fig. 8). We next showed that increasing GRF4 abundance improves NUE and grain yield of the high-yielding *sd1*-containing *indica* variety 9311. As for NJ6-*sd1-GRF4^{gr2}* (Fig. 3a), the 9311-*GRF4^{gr2}* isogenic line is not detectably changed with respect to the *sd1*-conferred semi-dwarf phenotype (Fig. 5a, b), but displays increased leaf and culm width (Extended Data Fig. 9a, b). However, the increased

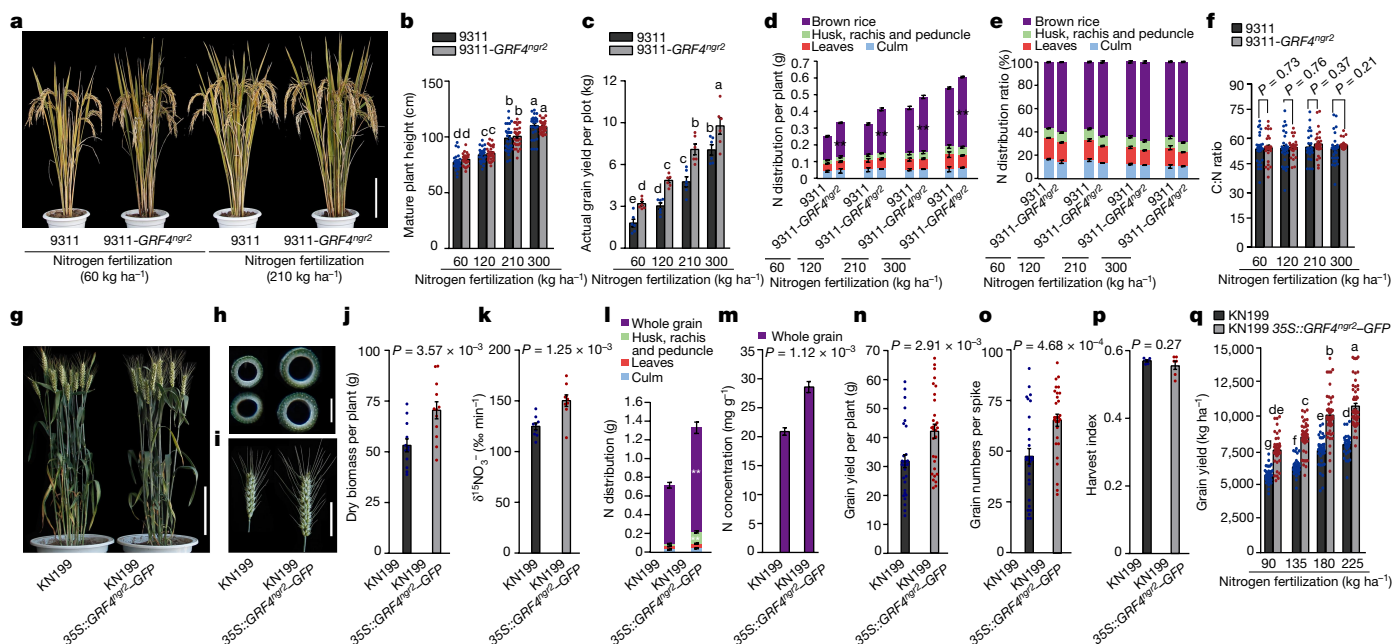


Fig. 5 | Increased GRF4 abundance boosts grain yield and NUE of rice and wheat GRVs without increasing mature plant height. **a**, Appearance of mature plants. Scale bar, 15 cm. **b**, Plant height. Data are mean \pm s.e.m. ($n = 40$). **c**, Grain yield. Data are mean \pm s.e.m. of six plots (each plot contained 220 plants) per line per nitrogen level. **d**, **e**, Absolute (**d**) and proportional (**e**) nitrogen distribution of plants shown in **b**. $**P < 0.05$, 9311-GRF4^{gr2} compared with 9311 by two-sided Student's *t*-test. **f**, C:N ratio of plants shown in **b**. Data in **d**–**f** are mean \pm s.e.m. ($n = 30$). **g**, Mature wheat plant morphology. Scale bar, 15 cm. **h**, Cross-section of the uppermost internode of (left) KN199 and (right) KN199 35S::GRF4^{gr2}-GFP wheat plants. Scale bar, 2 mm. **i**, Spike length. Scale bar,

5 cm. **a**, **g**–**i**, Pictures are representative of three experiments performed independently with similar results. **j**, Biomass accumulation. Data are mean \pm s.e.m. ($n = 12$). **k**, $^{15}\text{NO}_3^-$ uptake rate. **l**, Nitrogen distribution. Data in **k**, **l** are mean \pm s.e.m. ($n = 9$). $**P < 0.05$, KN199 35S::GRF4^{gr2}-GFP compared with KN199. **m**, Dry-weight nitrogen concentrations. Data are mean \pm s.e.m. ($n = 20$). **n**, Grain yield. **o**, Grain number. Data in **n**, **o** are mean \pm s.e.m. ($n = 30$). **p**, Harvest index. Data are mean \pm s.e.m. ($n = 6$). **q**, Overall grain yield. Data are mean \pm s.e.m. ($n = 60$). **f**, **j**, **k**, **m**–**p**, *P* values are from two-sided Student's *t*-tests. **b**, **c**, **q**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test.

$^{15}\text{NH}_4^+$ and $^{15}\text{NO}_3^-$ uptake conferred by GRF4^{gr2} (Extended Data Fig. 9c, d) enhances 9311 grain yield and NUE. Grain yield per plot was increased in 9311-GRF4^{gr2} (versus 9311) at both high and low levels of nitrogen supply (Fig. 5c), owing to increases in both grain number and grain weight^{18–20} (Extended Data Fig. 9e, f). Harvest index was relatively unaffected (Extended Data Fig. 9g), presumably because biomass increases (Extended Data Fig. 9h) balance out increases in grain yield (Fig. 5c). Although total nitrogen in above-ground parts of 9311-GRF4^{gr2} was greater than in 9311 (Fig. 5d), the distribution ratio of nitrogen allocated to grain (versus vegetative organs) was not significantly increased (Fig. 5e) and the C:N ratio was not detectably affected (Fig. 5f). Thus, the increased GRF4 abundance conferred by GRF4^{gr2} partially disconnects GA regulation of stem elongation (plant height) from nitrogen metabolic regulation. GRF4-promoted biomass increases are reflected primarily in increased leaf and culm widths rather than height.

Chinese *japonica* rice GRV semi-dwarfism is conferred by a mutant variant (*dep1-1*) of the γ subunit³² that reduces vegetative growth nitrogen response and increases NUE²². We found that increasing GRF4 abundance (GRF4-GFP in transgenic WYJ7-*dep1-1*²² plants expressing 35S::GRF4^{gr2}-GFP) did not suppress *dep1-1*-conferred semi-dwarfism (Extended Data Fig. 10a), but increased both $^{15}\text{NH}_4^+$ and $^{15}\text{NO}_3^-$ uptake rates (Extended Data Fig. 10b–d). In addition, although plant height, heading date and tiller numbers per plant in response to different nitrogen supply rates were unaffected (Extended Data Fig. 10e–g), overexpression of GRF4^{gr2}-GFP increased both grain number (in low nitrogen; Extended Data Fig. 10h) and grain yield (Extended Data Fig. 10i) of WYJ7-*dep1-1*. Nutrient assimilation and grain yield of rice GRVs can thus be increased by higher GRF4 abundance, particularly at low nitrogen fertilization levels, without simultaneously causing yield-reducing plant height increases.

Finally, the semi-dwarfism of high-yielding Chinese wheat GRV KN199 is conferred by the mutant *Rht-B1b* allele^{5,6}. As in rice, transgenic expression of 35S::GRF4^{gr2}-GFP did not increase KN199 plant

height (Fig. 5g), but did increase culm diameter and wall thickness (Fig. 5h), spike length (Fig. 5i) and biomass accumulation (Fig. 5j). In addition, 35S::GRF4^{gr2}-GFP increased the $^{15}\text{NO}_3^-$ -uptake rate of KN199 (Fig. 5k), total nitrogen in above-ground plant parts (Fig. 5l) and nitrogen concentration in de-husked grain (Fig. 5m). 35S::GRF4^{gr2}-GFP also boosted KN199 yield (Fig. 5n) by increasing grain numbers per spike (Fig. 5o), without affecting harvest index (Fig. 5p). Moreover, the improvement of grain yield conferred on KN199 by 35S::GRF4^{gr2}-GFP at low nitrogen supply shows that increased GRF4 abundance enhances both grain yield and NUE of wheat GRVs (Fig. 5q), without affecting the beneficial GRV semi-dwarfism. Indeed, the increased culm width and wall thickness conferred by 35S::GRF4^{gr2}-GFP (Fig. 5h) is likely to enhance the stem robustness conferred by mutant *Rht* alleles, thus further reducing lodging yield loss. In conclusion, increased GRF4 abundance increases grain yields of rice and wheat GRVs grown in moderate nitrogen conditions.

Discussion

We here report new advances in fundamental plant science and strategic plant breeding. First, the GRF4–DELLA interaction integrates plant growth and metabolic regulation. GRF4 is a transcriptional regulator of multiple nitrogen-metabolism genes that, because it is itself nitrogen regulated, probably confers homeostatic coordination of plant nitrogen metabolism. Notably, nitrogen-regulated GRF4 also coordinates carbon metabolism and growth, and is thus likely to confer broader-range integrative homeostatic control. Although long thought to exist, the identities of such broad-range growth and metabolic integrators were previously unknown. Furthermore, GRF4 activity is balanced by an antagonistic regulatory relationship with the DELLA growth repressor. Essentially, physical DELLA–GRF4–GIF1 interactions enable DELLA to inhibit activation of target gene promoters by GRF4–GIF1, and the balance between opposing GRF4 and DELLA activities thus enhances coordinated regulation of plant growth and metabolism.

Second, increasing the abundance of GRF4 in GRVs tips the GRF4–DELLA balance to favour GRF4, conferring increases in carbon and nitrogen assimilation, biomass, leaf and stem width, but having little effect on plant height³³. The practical plant breeding consequence of this is that it enables enhanced GRV nutrient assimilation without the loss of the beneficial semi-dwarfism conferred by DELLA accumulation. GRV NUE can thus be improved, without the yield penalties of increased lodging. Genetic variation of *GRF4* (and orthologues) should now become a major target for breeders in enhancing crop yield and nutrient-use efficiency. Such enhancements will enable future green revolutions, sustainably increasing yield, yet reducing environmentally degrading agricultural nitrogen use.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0415-5>.

Received: 9 February 2018; Accepted: 5 July 2018;

Published online 15 August 2018.

- Khush, G. S. Green revolution: preparing for the 21st century. *Genome* **42**, 646–655 (1999).
- Pingali, P. L. Green revolution: impacts, limits, and the path ahead. *Proc. Natl Acad. Sci. USA* **109**, 12302–12308 (2012).
- Evenson, R. E. & Gollin, D. Assessing the impact of the green revolution, 1960 to 2000. *Science* **300**, 758–762 (2003).
- Hedden, P. The genes of the green revolution. *Trends Genet.* **19**, 5–9 (2003).
- Peng, J. et al. ‘Green revolution’ genes encode mutant gibberellin response modulators. *Nature* **400**, 256–261 (1999).
- Zhang, C., Gao, L., Sun, J., Jia, J. & Ren, Z. Haplotype variation of green revolution gene *Rht-D1* during wheat domestication and improvement. *J. Integr. Plant Biol.* **56**, 774–780 (2014).
- Sasaki, A. et al. Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* **416**, 701–702 (2002).
- Spielmeyer, W., Ellis, M. H. & Chandler, P. M. Semidwarf (*sd-1*), “green revolution” rice, contains a defective gibberellin 20-oxidase gene. *Proc. Natl Acad. Sci. USA* **99**, 9043–9048 (2002).
- Harberd, N. P., Belfield, E. & Yasumura, Y. The angiosperm gibberellin–GID1–DELLA growth regulatory mechanism: how an “inhibitor of an inhibitor” enables flexible response to fluctuating environments. *Plant Cell* **21**, 1328–1339 (2009).
- Xu, H., Liu, Q., Yao, T. & Fu, X. Shedding light on integrative GA signaling. *Curr. Opin. Plant Biol.* **21**, 89–95 (2014).
- Itoh, H., Ueguchi-Tanaka, M., Sato, Y., Ashikari, M. & Matsuoka, M. The gibberellin signaling pathway is regulated by the appearance and disappearance of SLENDER RICE1 in nuclei. *Plant Cell* **14**, 57–70 (2002).
- Asano, K. et al. Artificial selection for a green revolution gene during japonica rice domestication. *Proc. Natl Acad. Sci. USA* **108**, 11034–11039 (2011).
- Gooding, M. J., Addisu, M., Uppal, R. K., Snape, J. W. & Jones, H. E. Effect of wheat dwarfing genes on nitrogen-use efficiency. *J. Agric. Sci.* **150**, 3–22 (2012).
- Li, B.-Z. et al. Molecular basis and regulation of ammonium transporter in rice. *Rice Sci.* **16**, 314–322 (2009).
- Hawkesford, M. J. Reducing the reliance on nitrogen fertilizer for wheat production. *J. Cereal Sci.* **59**, 276–283 (2014).
- Zhao, X. et al. Nitrogen runoff dominates water nitrogen pollution from rice-wheat rotation in the Taihu Lake region of China. *Agric. Ecosyst. Environ.* **156**, 1–11 (2012).
- Conway, G. *One Billion Hungry. Can We Feed the World?* (Cornell Univ. Press, Ithaca, 2012).
- Che, R. et al. Control of grain size and rice yield by *GL2*-mediated brassinosteroid responses. *Nat. Plants* **2**, 15195 (2015).
- Duan, P. et al. Regulation of *OsGRF4* by *OsmiR396* controls grain size and yield in rice. *Nat. Plants* **2**, 15203 (2015).
- Hu, J. et al. A rare allele of *GS2* enhances grain size and grain yield in rice. *Mol. Plant* **8**, 1455–1465 (2015).
- Ma, X. et al. A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol. Plant* **8**, 1274–1284 (2015).
- Sun, H. et al. Heterotrimeric G proteins regulate nitrogen-use efficiency in rice. *Nat. Genet.* **46**, 652–656 (2014).
- Somers, D. A., Kuo, T. M., Kleinhofs, A., Warner, R. L. & Oaks, A. Synthesis and degradation of barley nitrate reductase. *Plant Physiol.* **72**, 949–952 (1983).
- Tabuchi, M., Abiko, T. & Yamaya, T. Assimilation of ammonium ions and reutilization of nitrogen in rice (*Oryza sativa* L.). *J. Exp. Bot.* **58**, 2319–2327 (2007).
- Peng, J. et al. The *Arabidopsis* *GAI* gene defines a signaling pathway that negatively regulates gibberellin responses. *Genes Dev.* **11**, 3194–3205 (1997).
- Nunes-Nesi, A., Fernie, A. R. & Stitt, M. Metabolic and signaling aspects underpinning the regulation of plant carbon nitrogen interactions. *Mol. Plant* **3**, 973–996 (2010).
- LeCain, D. R., Morgan, J. A. & Zerbi, G. Leaf anatomy and gas exchange in nearly isogenic semidwarf and tall winter wheat. *Crop Sci.* **29**, 1246–1251 (1989).
- Morgan, J. A., LeCain, D. R. & Wells, R. Semidwarfing genes concentrate photosynthetic machinery and affect leaf gas exchange of wheat. *Crop Sci.* **30**, 602–608 (1990).
- Fabian, T., Lorbiecke, R., Umeda, M. & Sauter, M. The cell cycle genes *cycA1;1* and *cdc20s-3* are coordinately regulated by gibberellin in planta. *Planta* **211**, 376–383 (2000).
- Sauter, M. Differential expression of a CAK (*cdc2*-activating kinase)-like protein kinase, cyclins and *cdc2* genes from rice during the cell cycle and in response to gibberellin. *Plant J.* **11**, 181–190 (1997).
- Yu, J. et al. *OsLG3* contributing to rice grain length and yield was mined by Ho-LAMap. *BMC Biol.* **15**, 28 (2017).
- Huang, X. et al. Natural variation at the *DEP1* locus enhances grain yield in rice. *Nat. Genet.* **41**, 494–497 (2009).
- Serrano-Mislata, A. et al. *DELLA* genes restrict inflorescence meristem function independently of plant height. *Nat. Plants* **3**, 749–754 (2017).

Acknowledgements We thank J. F. Ma for comments on this manuscript. This research was supported by grants from the National Key Research and Development Program of China (2016YFD0100401, 2016YFD0100706 and 2016YFD0100901), National Natural Science Foundation of China (91635302), Chinese Academy of Sciences (XDA08010101) and by the Biological and Biotechnological Sciences Research Council (UK) ‘Newton Fund’ Rice Research Initiative grant BB/M011224/1.

Reviewer information *Nature* thanks B. Hirel, M. Matsuoka and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.L. performed most of the experiments; Y.Ti. and S.L. conducted QTL analysis; S.L., J.Z. and K.W. constructed near isogenic lines; S.L., Y.Y. and Q.L. performed field experiments; Y.To., M.H. and H.L. characterized the phenotypes of transgenic wheat plants; J.Y. performed haplotype analysis; N.P.H. and X.F. designed experiments; N.P.H. and X.F. wrote the manuscript. All authors discussed the results and contributed to the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0415-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0415-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to X.F.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Plant materials and field growth conditions. Details of rice germplasm used for positional cloning and haplotype analysis have been described elsewhere^{22,31,34}. QTL analysis and map-based cloning were performed using BC₁F₂, BC₂F₂ and BC₃F₂ populations derived from a cross between selected variety NM73 and *indica* variety NJ6 (the recurrent parent). Near isogenic line (NIL) plants carrying differing combinations of the *qngr2* and *sd1* alleles were bred by crossing NM73 × NJ6 and NM73 × 9311 F₁ six times with NJ6, NJ6-*sd1* and 9311 as recurrent parents. Field-grown NILs and transgenic rice plants were raised in standard paddy conditions with an interplant spacing of 20 cm at the Institute of Genetics and Developmental Biology experimental station sites located in Lingshui (Hainan Province), Hefei (Anhui Province) and Beijing as previously described^{22,32}. Field-grown wheat plants (Chinese wheat GRV KN199 and transgenic derivatives) were planted during the winter planting season at the Experimental Station of the Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences (Shijiazhuang, Hebei province).

Hydroponic culture conditions. Hydroponic culture conditions were modified from previously published work³⁵. Seeds were disinfected in 20% sodium hypochlorite solution for 30 min, thoroughly washed with deionized water, and then germinated in moist Perlite. Seven-day-old seedlings were then selected and transplanted to PVC pots containing 40 l + nitrogen nutrient solution (1.25 mM NH₄NO₃, 0.5 mM NaH₂PO₄·2H₂O, 0.75 mM K₂SO₄, 1 mM CaCl₂, 1.667 mM MgSO₄·7H₂O, 40 μM Fe-EDTA (Na), 19 μM H₃BO₃, 9.1 μM MnSO₄·H₂O, 0.15 μM ZnSO₄·7H₂O, 0.16 μM CuSO₄ and 0.52 μM (NH₄)₃Mo₇O₂₄·4H₂O, pH 5.5). The compositions of nutrient solutions containing different levels of supplied nitrogen were as follows: 1N, 1.25 mM NH₄NO₃; 0.6N, 0.75 mM NH₄NO₃; 0.3N, 0.375 mM NH₄NO₃; 0.15N, 0.1875 mM NH₄NO₃. All nutrient solutions were changed twice per week, pH was adjusted to 5.5 every day. The temperature was maintained at 30 °C day and 22 °C night, and the relative humidity was 70%.

Positional cloning of *qNGR2*. The map-based cloning of *qngr2* was based on 1,849 BC₂F₂ and 3,124 BC₃F₂ populations derived from the backcross between the selected variety NM73 and the *indica* rice variety NJ6 (with NJ6 as the recurrent parent). Primer sequences used for map-based cloning and genotyping assays are given in Supplementary Table 5.

Transgene constructs. The *GRF4*^{NGR2} mRNA-encoding sequence (together with intron sequences) was amplified from NJ6. The *GRF4*^{NGR2} mRNA-coding sequence (together with introns and/or promoter regions lying approximately 3-kb upstream of the transcription start site) was amplified from NM73. These amplified genomic DNA fragments were then inserted into the *Actin::nos*³⁶ and *CAMBIA2300* (Cambia, <http://www.cambia.org/>) vectors to respectively generate the *Actin::GRF4*^{NGR2} and *GRF4*^{NGR2} constructs. A full-length *GRF4*^{NGR2} cDNA was introduced into the 35S::GFP-*nos*²² and 35S::Flag-*nos*³⁴ vectors to generate the 35S::GRF4^{NGR2}-GFP and 35S::Flag-GRF4^{NGR2} constructs. A 300-bp *GRF4*^{NGR2} cDNA fragment was amplified and used to construct the *Actin::RNAi-GRF4* transgene, as described elsewhere³². gRNA constructs required for construction of the CRISPR-Cas9-generated *GRF4* loss of function allele (*grf4*) in the WYJ7 genetic background were made as described elsewhere^{21,34}. Transgenic rice and wheat plants were generated by *Agrobacterium*-mediated transformation as described elsewhere³². Relevant primer sequences are given in Supplementary Table 6.

RT-qPCR. Total RNAs were extracted from different organs of three-week-old rice plants under hydroponic conditions using the TRIzol reagent (Invitrogen), and then treated with RNase-free DNase I (Invitrogen) according to the manufacturer's protocol. Full-length cDNA was then reverse transcribed using a cDNA synthesis kit (TRANSGEN, AE311). qPCR was performed according to the manufacturer's instructions (TRANSGEN, AQ101), using three independent RNA preparations as biological replicates. Rice *Actin2* gene transcripts were used as a reference. The relevant primer sequences are given in Supplementary Table 7.

Bimolecular fluorescence complementation (BiFC) assays. The full-length cDNAs corresponding to the *SLR1*, *GIF1*, *GIF2*, *GIF3*, *GRF1*, *GRF2*, *GRF3*, *GRF4*, *GRF5*, *GRF6*, *GRF7*, *GRF8*, *GRF9*, *GRF10*, *GRF11* and *GRF12* genes, along with both deleted and non-deleted versions of *GRF4* cDNA were amplified from NJ6. The resultant amplicons were inserted into the pSY-735-35S-cYFP-HA or pSY-736-35S-nYFP-EE vectors³⁷ to generate fusion constructs. Co-transfection of constructs (for example, those encoding nYFP-GRF4 and cYFP-SLR1) into tobacco leaf epidermal cells by *Agrobacterium*-mediated infiltration enabled testing for protein-protein interactions. After 48 h incubation in the dark, the YFP signal was examined and photographed using a confocal microscope (Zeiss LSM710). Each BiFC assay was repeated at least three times. Relevant primer sequences are given in Supplementary Table 6.

Co-immunoprecipitation and western blotting. Full-length *GRF4*, *GIF1* and *SLR1* cDNAs were amplified, and then inserted into either the pUC-35S-HA-RBS or the pUC-35S-Flag-RBS vector as previously described³⁸. *A. thaliana* protoplasts were transfected with 100 μg of plasmid and then incubated overnight in low light intensity conditions. Total protein was then extracted from harvested

protoplasts by treating with 50 mM HEPES (pH 7.5), 150 mM KCl, 1 mM EDTA (pH8), 0.3% Triton X-100, 1 mM DTT with added proteinase inhibitor cocktail (Roche LifeScience). Lysates were incubated with magnetic beads conjugated with an anti-DDDDK-tag antibody (MBL, M185-11) at 4 °C for at least 4 h. The magnetic beads were then rinsed six times with the extraction buffer and eluted with 3 × Flag peptide (Sigma-Aldrich, F4709). Immunoprecipitates were electrophoretically separated by SDS-PAGE and transferred to a nitrocellulose membrane (GE Healthcare). Proteins were detected by immunoblot using the antibodies anti-Flag (Sigma, F1804) and anti-HA (MBL, M180-7). In addition, the GRF4, SLR1, Lhca1, Lhca3, Lhca4, Lhcb2, PsdA and PsdE proteins were detected by probing the membrane with anti-GRF4 antibodies (Abmart), anti-SLR1 antibodies (ABclonal Technology), anti-Lhca1 antibodies (Agrisera, AS01005), anti-Lhca3 antibodies (Agrisera, AS01007), anti-Lhca4 antibodies (Agrisera, AS01008), anti-Lhcb2 antibodies (Agrisera, AS01003), anti-PsdA antibodies (Agrisera, AS09461) and anti-PsdE antibodies (Agrisera, AS08324A), respectively. Uncropped blots are shown in Supplementary Fig. 1. Relevant primer sequences are given in Supplementary Table 6.

EMSA assays. EMSA was performed as previously described with minor modifications³⁹. Full-length *GIF1* and *SLR1* cDNAs were amplified and cloned into the pCold-TF vector (Takara). His-GRF1 and His-SLR1 recombinant proteins were purified using Ni-NTA agarose (QIAGEN, 30210), following the manufacturer's instructions. GST and GST-GRF4 recombinant proteins were expressed in the *Escherichia coli* BL21 (DE3) strain and then purified using Glutathione Sepharose 4B beads (GE Healthcare, 17-0756-01). DNA probes (42 bp) were artificially amplified and labelled using a biotin label kit (Biosune). DNA gel shift assays were performed using the LightShift Chemiluminescent EMSA kit (Thermo Fisher Scientific, 20148). Relevant primer sequences are given in Supplementary Table 8.

RNA-seq analysis. Total RNAs were extracted from three-week-old rice plants grown under high nitrogen conditions (1.25 mM NH₄NO₃) using the QIAGEN RNeasy plant mini kit (QIAGEN, 74904) following the manufacturer's instructions. Three replicate RNA-seq libraries were prepared from NJ6, NJ6-*sd1* and NJ6-GRF4^{NGR2} plants. A total of the nine libraries were sequenced separately using the BGISEQ-500 sequencer. For each RNA sample, the NIL plants were collected from three replicates and pooled together after RNA extraction. Raw sequencing reads were cleaned by removing adaptor sequences, reads containing poly-N sequences, and low-quality reads. Approximately 24,006,405 clean reads were mapped to the Nipponbare reference genome using HISAT40/Bowtie241 tools. After data were mapped, normalization was performed and then FPKM (fragments per kilobase per million mapped reads) was calculated using RESM software⁴². As previously described⁴³, a false discovery rate (FDR) < 0.01 and absolute value of log₂ ratio ≥ 2 were used to identify differentially expressed genes in NJ6-*sd1* versus NJ6 and NJ6-GRF4^{NGR2} versus NJ6 samples. Comparisons of the three individual replicate FPKM values of the genes involved in the coordinated regulation of plant growth and nitrogen, and carbon metabolism are given in Supplementary Table 3.

ChIP-seq and ChIP-qPCR assays. ChIP assays were performed as previously described with minor modifications⁴⁴. Approximately 2 g of two-week-old seedlings of transgenic 35S::Flag-GRF4^{NGR2} rice plants grown under the high nitrogen (1.25 mM NH₄NO₃) conditions were fixed with 1% (v/v) formaldehyde under vacuum for 15 min at 20–25 °C, and then homogenized in liquid nitrogen. After isolation and lysing of nuclei, the chromatin complexes were isolated and ultrasonically fragmented into fragments with an average size of approximately 500 bp. Immunoprecipitations were performed with anti-Flag antibodies (Sigma, F1804) overnight at 4 °C. The precipitated DNA was recovered and dissolved in water and stored at –80 °C. Illumina sequencing libraries were constructed according to the manufacturer's instructions, and then sequenced on the BGISEQ-500 platform. Sequencing reads were mapped to the Nipponbare reference genome using SOAP aligner/soap245. The peak summits were used to define the peak location types on the genome, and motif search and classification were performed as previously described⁴⁶. In addition, the precipitated DNA samples served as template for RT-qPCR. Relevant primer sequences are given in Supplementary Table 9.

FRET assay. Cauliflower mosaic virus 35S promoter-driven fusion constructs with C-terminal tagging of CFP or YFP were created to generate the donor vector 35S::GIF1-CFP and the acceptor vector 35S::GRF4-YFP. Donor and acceptor vectors, with or without a 35S::SLR1 vector and/or GA (GA₃), were co-transformed into tobacco leaf epidermis cells by *Agrobacterium*-mediated infiltration to provide the FRET signal. Transformation with 35S::GIF1-CFP vector only provided the donor signal, and transformation with 35S::GRF4-YFP vector only provided the acceptor signal. The FRET signal was detected and photographed using a confocal microscope (Zeiss LSM710). Relevant primer sequences are given in Supplementary Table 6.

In vitro transient transactivation assays. Approximately 2-kb DNA promoter fragments from each of *AMT1.1*, *AMT1.2*, *NRT1.1B*, *NRT2.3a*, *NPF2.4*, *GS1.2*, *GS2*, *NADH-GOGAT2*, *Fd-GOGAT*, *NIA1*, *NIA3*, *NiR1*, *PsS1*, *TPS1*, *SWEET11*, *cycA1*, *cdc2-3*, or *GRF4* were amplified from NJ6, and then subcloned into a pUC19 vector

containing the firefly luciferase reporter gene driven by the 35S minimal TATA box and 5 × GAL4 binding elements, thus generating reporter plasmids containing specific promoters fused to luciferase. The full-length *GRF4* cDNA was amplified and fused to sequence encoding GAL4BD, thus generating the effector plasmid *RTBD-GRF4*. Transient transactivation assays were performed using rice protoplasts as described elsewhere⁴⁷. The Dual-Luciferase Reporter Assay System (Promega, E1960) was used to perform the luciferase activity assay, with the Renilla luciferase gene as an internal control. Relevant primer sequences are given in Supplementary Table 6.

Determination of plant carbon and nitrogen concentrations. Samples from various plant organs were dried in an oven at 80 °C for 72 h. After tissue homogenization, carbon and nitrogen concentrations were determined using an elemental analyser (IsoPrime100; Elementar). All experiments were conducted with at least three replicates.

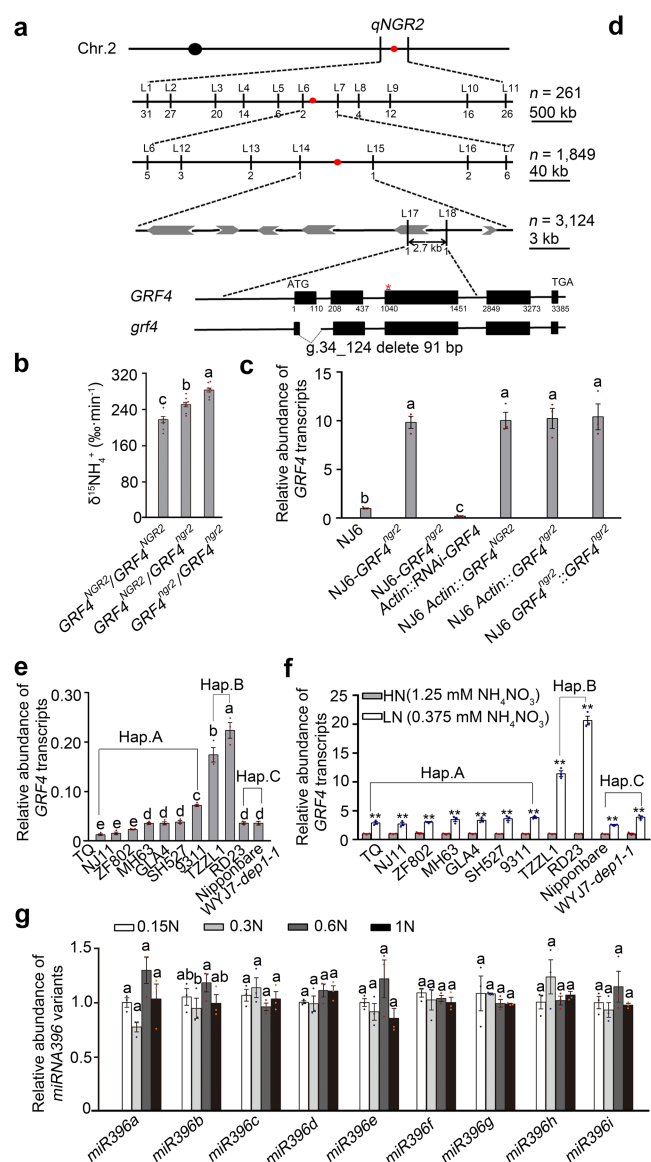
¹⁵N uptake analysis. After growth in hydroponic culture for 4 weeks, rice root ¹⁵NO₃⁻ and ¹⁵NH₄⁺ influx measurements were as described elsewhere^{48,49}. Roots and shoots were separated and stored at -70 °C before freeze drying. Roots and shoots were dried overnight at 80 °C, and the ¹⁵N content was measured using the IsoPrime 100 (Elementar, Germany).

Determination of glutamine synthase and nitrate reductase activities. Glutamine synthase and nitrate reductase activities were determined with the Glutamine Synthetase Kit (Solarbio LIFE SCIENCES, BC0910) and the Nitrate Reductase Kit (Solarbio LIFE SCIENCES, BC0080) following the manufacturer's instructions.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

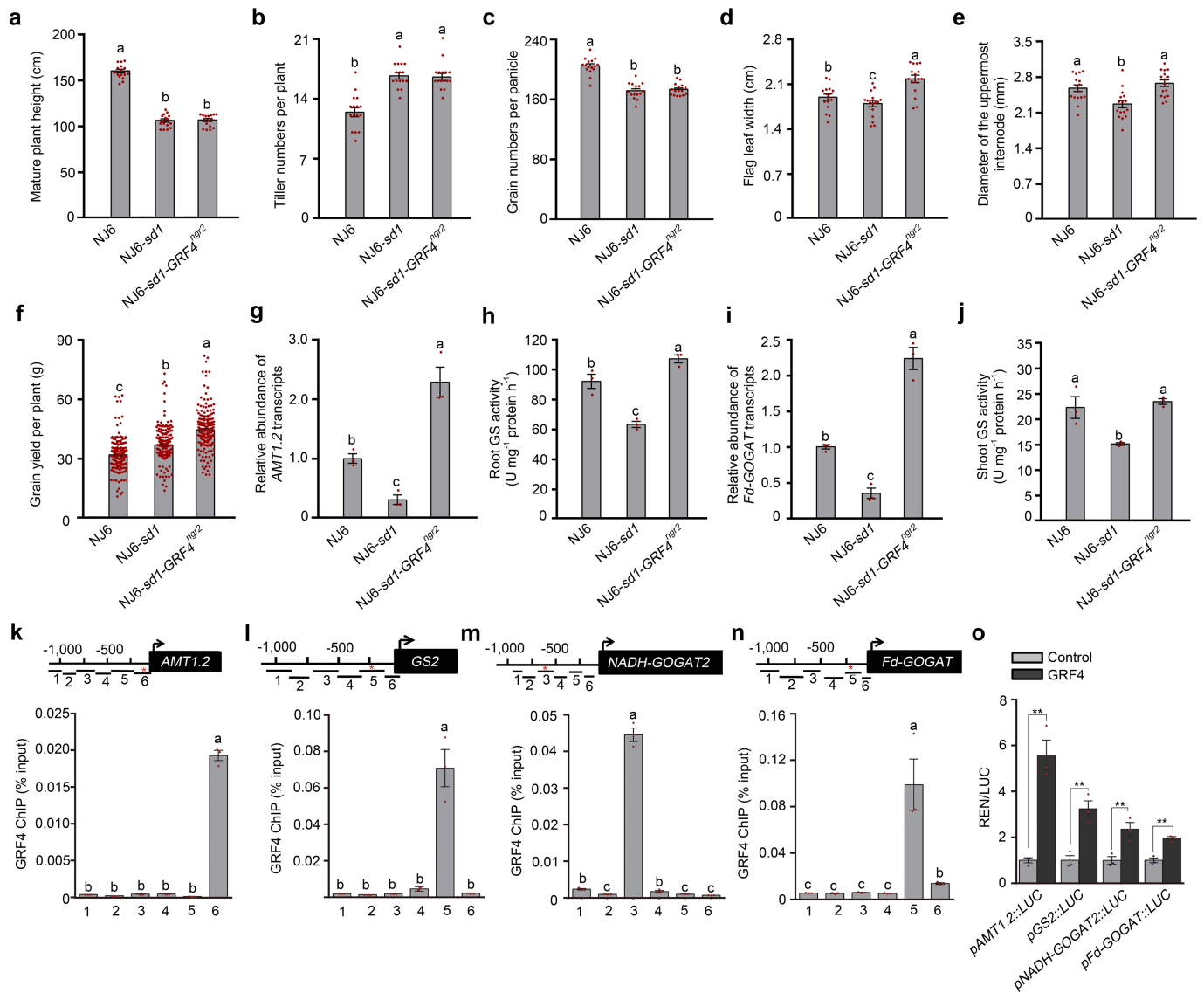
Data availability. Sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE114287. Source Data (including uncropped western blots and RNA-seq data) generated and/or analysed in the current study and Supplementary Information are provided with the online version of the paper. All other data are available from the corresponding author upon reasonable request.

34. Wang, S. et al. Non-canonical regulation of SPL transcription factors by a human OTUB1-like deubiquitinase defines a new plant type rice associated with higher grain yield. *Cell Res.* **27**, 1142–1156 (2017).
35. Liu, W.-J., Zhu, Y.-G., Smith, F. A. & Smith, S. E. Do phosphorus nutrition and iron plaque alter arsenate (As) uptake by rice seedlings in hydroponic culture? *New Phytol.* **162**, 481–488 (2004).
36. Wang, S. et al. Control of grain size, shape and quality by *OsSPL16* in rice. *Nat. Genet.* **44**, 950–954 (2012).
37. Bracha-Drori, K. et al. Detection of protein–protein interactions in plants using bimolecular fluorescence complementation. *Plant J.* **40**, 419–427 (2004).
38. Chen, H. et al. Firefly luciferase complementation imaging assay for protein–protein interactions in plants. *Plant Physiol.* **146**, 368–376 (2008).
39. Chen, L. et al. *OsMADS57* together with *OsTB1* coordinates transcription of its target *OsWRKY94* and *D14* to switch its organogenesis to defense for cold adaptation in rice. *New Phytol.* **218**, 219–231 (2018).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
42. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
43. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284 (2001).
44. O'Geen, H., Fietze, S. & Farnham, P. J. Using ChIP-seq technology to identify targets of zinc finger transcription factors. *Methods Mol. Biol.* **649**, 437–455 (2010).
45. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
46. Lu, Z. et al. Genome-wide binding analysis of the transcription activator ideal plant architecture1 reveals a complex network regulating rice plant architecture. *Plant Cell* **25**, 3743–3759 (2013).
47. Wang, S. et al. The *OsSPL16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* **47**, 949–954 (2015).
48. Ho, C. H., Lin, S. H., Hu, H. C. & Tsay, Y. F. CHL1 functions as a nitrate sensor in plants. *Cell* **138**, 1184–1194 (2009).
49. Loqué, D. et al. Additive contribution of *AMT1;1* and *AMT1;3* to high-affinity ammonium uptake across the plasma membrane of nitrogen-deficient *Arabidopsis* roots. *Plant J.* **48**, 522–534 (2006).



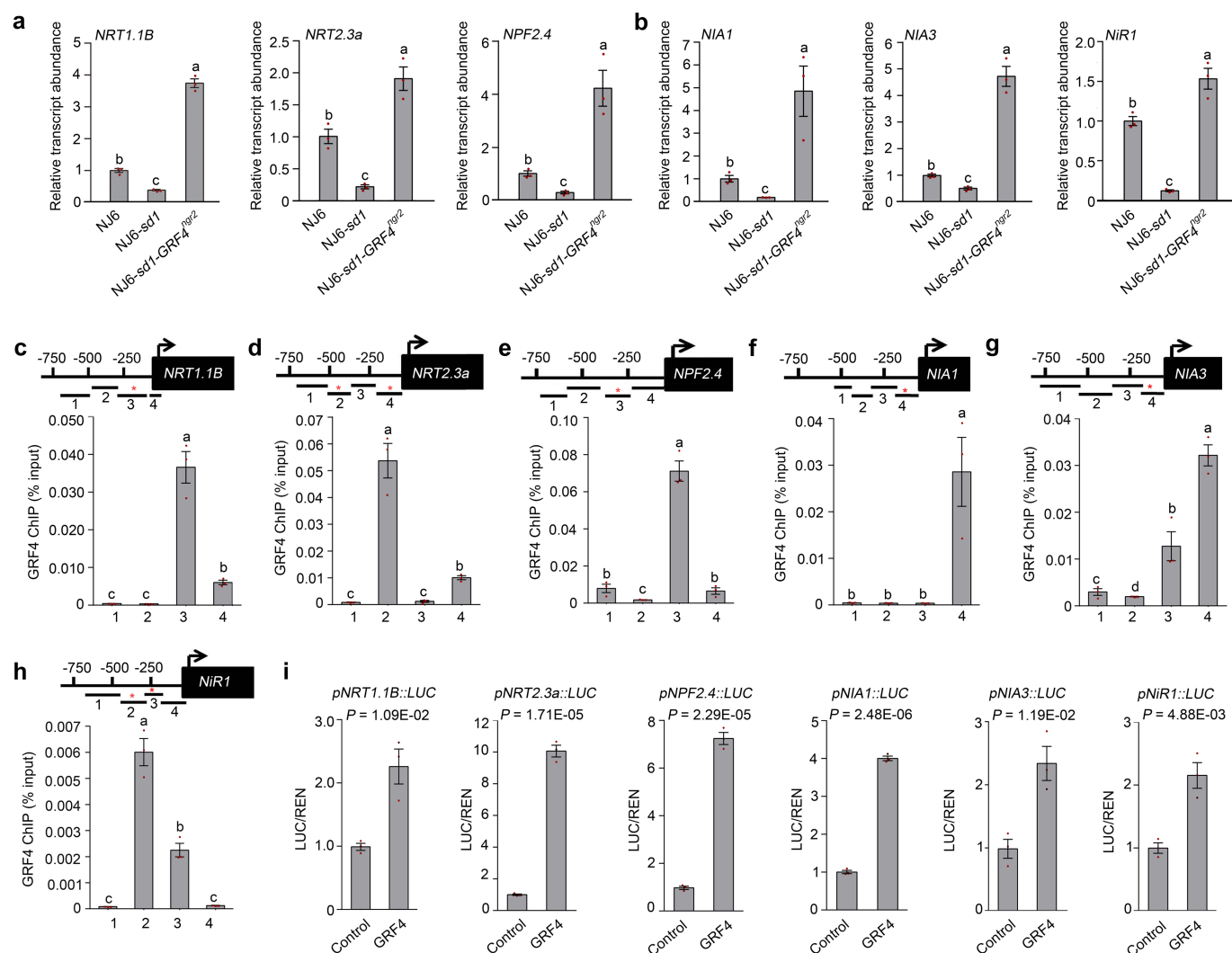
Extended Data Fig. 1 | Allelic variation at the *GRF4* locus affects *GRF4* mRNA abundance and root $^{15}\text{NH}_4^+$ uptake. **a**, Positional cloning indicates the equivalence of *GRF4* with *qNGR2* (nitrogen-mediated growth response 2). Successive maps show progressive narrowing of focus of *qNGR2* (red dot, using recombination break points and linked DNA markers) to an approximately 2.7-kb region on chromosome 2 flanked by molecular markers L17 and L18 and overlapping candidate gene LOC_02g47280 (also known as *GRF4*). The start ATG (nucleotide 1) and close TGA (nucleotide 3385) of *GRF4* are shown, together with the protein-coding DNA sequence (thick black bars). The target site for miR396 is indicated by an asterisk. The structure of a CRISPR-Cas9-generated *grf4* mutant 91-bp deletion allele spanning parts of exon 1 and intron 1 is shown. **b**, $^{15}\text{NH}_4^+$ uptake rates of roots of BC₂F₂ progeny (derived from a NJ6 × NM73 cross) homozygous or heterozygous for *GRF4*^{NGR2} or *GRF4*^{gr2} grown in high nitrogen supply (1.25 mM NH₄NO₃). Data are mean ± s.e.m. (*n* = 9). Different letters denote significant differences (*P* < 0.05) from a Duncan's multiple range test. **c**, *GRF4* mRNA abundance in plants (genotypes as shown) relative to the abundance in NJ6 (set to one). Data are mean ± s.e.m. (*n* = 3). Different letters denote significant differences (*P* < 0.05) from a Duncan's multiple range test. **d**, Natural

variety *GRF4* allelic variation. Nucleotide position relative to the *GRF4* start ATG is shown in **a**. SNPs shared between varieties NM73, RD23 and TZZL1 are highlighted. Sequences representative of *GRF4* promoter haplotypes A, B and C (see main text) are shown. **e**, *GRF4* mRNA abundance in various rice varieties under the high nitrogen conditions (1.25 mM NH₄NO₃), *GRF4* promoter haplotypes are indicated. Abundance data are all relative to the abundance of rice *Actin2* mRNA. Data are mean ± s.e.m. (*n* = 3). Different letters denote significant differences (*P* < 0.05) from a Duncan's multiple range test. **f**, Comparisons of *GRF4* mRNA abundance in selected rice varieties grown in between high (HN, 1.25 mM NH₄NO₃) and low (LN, 0.375 mM NH₄NO₃) nitrogen conditions. Data are mean ± s.e.m. (*n* = 3). Abundance data are all relative to the high nitrogen condition (set to one). ***P* < 0.05 compared to high nitrogen in a two-sided Student's *t*-test. **g**, Relative abundances of rice miR396 family members in NJ6 plants grown at different levels of nitrogen supply (0.15N, 0.1875 mM NH₄NO₃; 0.3N, 0.375 mM NH₄NO₃; 0.6N, 0.75 mM NH₄NO₃; 1N, 1.25 mM NH₄NO₃), shown relative to the abundance in plants grown in 1N conditions (set to one). Data are mean ± s.e.m. (*n* = 3). Different letters denote significant differences (*P* < 0.05) from a Duncan's multiple range test.



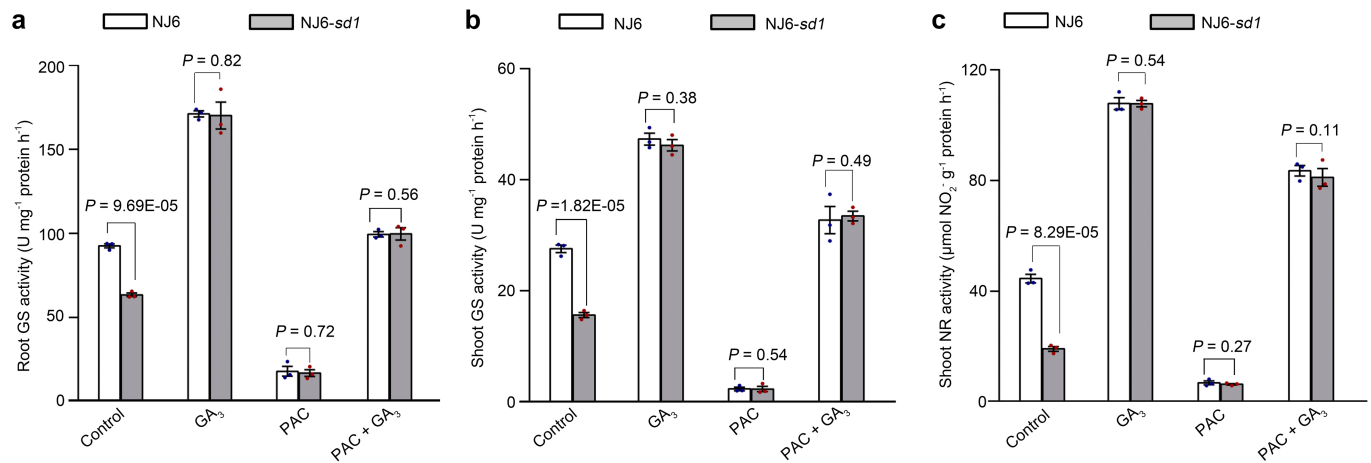
Extended Data Fig. 2 | Comparisons NJ6, NJ6-*sd1* and NJ6-*sd1-GRF4*^{ngr2} isogenic line traits reveals that GRF4 regulates expression of NH_4^+ -metabolism genes. a, Mature plant height. Data are mean \pm s.e.m. ($n = 16$). **b**, The number of tillers per plant. Data are mean \pm s.e.m. ($n = 16$). **c**, The number of grains per panicle. Data are mean \pm s.e.m. ($n = 16$). **d**, Flag-leaf width. Data are mean \pm s.e.m. ($n = 16$). **e**, Culm (stem) width expressed as diameter of the uppermost internode. Data are mean \pm s.e.m. ($n = 16$). **f**, Grain yield per plant. Data are mean \pm s.e.m. ($n = 220$). **g**, Relative root abundance of *AMT1.2* mRNA in NILs, genotypes as indicated. Abundances shown are relative to NJ6 plants (set to 1). Data are mean \pm s.e.m. ($n = 3$). **h**, Root glutamine synthase (GS) activities. Data are mean \pm s.e.m. ($n = 3$). **i**, Relative shoot abundance of *Fd-GOGAT* mRNA. Abundances shown are relative to NJ6 plants (set to 1).

Data are mean \pm s.e.m. ($n = 3$). **j**, Shoot glutamine synthase (GS) activities. Data are mean \pm s.e.m. ($n = 3$). **k–n**, Flag-GRF4-mediated ChIP-PCR enrichment (relative to input) of GCGG-containing promoter fragments (marked with an asterisk) from *AMT1.2*, *GS2*, *NADH-GOGAT2* and *Fd-GOGAT* promoters. Diagrams depict putative *AMT1.2*, *GS2*, *NADH-GOGAT2* and *Fd-GOGAT* promoters and fragments (1–6). Data are mean \pm s.e.m. ($n = 3$; panels **k–n**). **a–n**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test. **o**, GRF4 activates *AMT1.2*, *GS2*, *NADH-GOGAT2* and *Fd-GOGAT* promoter-luciferase fusion constructs in transient transactivation assays. Data are mean \pm s.e.m. ($n = 3$). ****** $P < 0.05$ compared to control group by two-sided Student's *t*-tests.



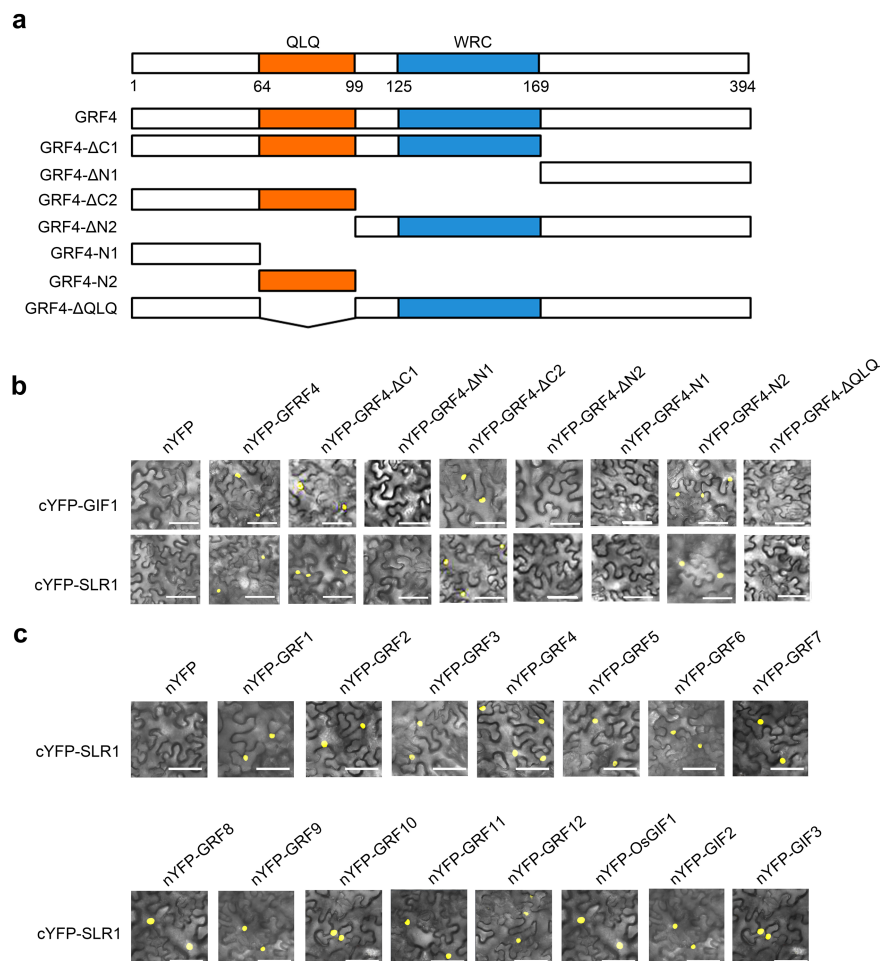
Extended Data Fig. 3 | GRF4 regulates expression of multiple NO_3^- metabolism genes. **a**, Relative abundance of *NRT1.1B*, *NRT2.3a* and *NPF2.4* mRNAs that encode NO_3^- uptake transporters. Abundances shown are relative to NJ6 (set to 1). Data are mean \pm s.e.m. ($n = 3$). **b**, Relative abundances of *NIA1*, *NIA3* and *NiR1* mRNAs that encode NO_3^- -assimilation enzymes. Abundances shown are relative to NJ6 (set to 1). Data are mean \pm s.e.m. ($n = 3$). **c-h**, Flag-GRF4-mediated ChIP-PCR enrichment (relative to input) of GCGG-containing fragments (marked

with asterisks) from promoters of *NRT1.1B* (c), *NRT2.3a* (d) and *NPF2.4* (e) genes that encode NO_3^- -uptake transporters and *NIA1* (f), *NIA3* (g) and *NiR1* (h) genes that encode NO_3^- -assimilation enzymes. Data are mean \pm s.e.m. ($n = 3$). **a-h**, Different letters denote significant differences ($P < 0.05$) from a Duncan's multiple range test. **i**, GRF4 activates *NRT1.1B*, *NRT2.3a*, *NPF2.4*, *NIA1*, *NIA3* and *NiR1* promoter-luciferase fusion constructs in transient transactivation assays. Data are mean \pm s.e.m. ($n = 3$) in all panels. P values are from a two-sided Student's t -test.



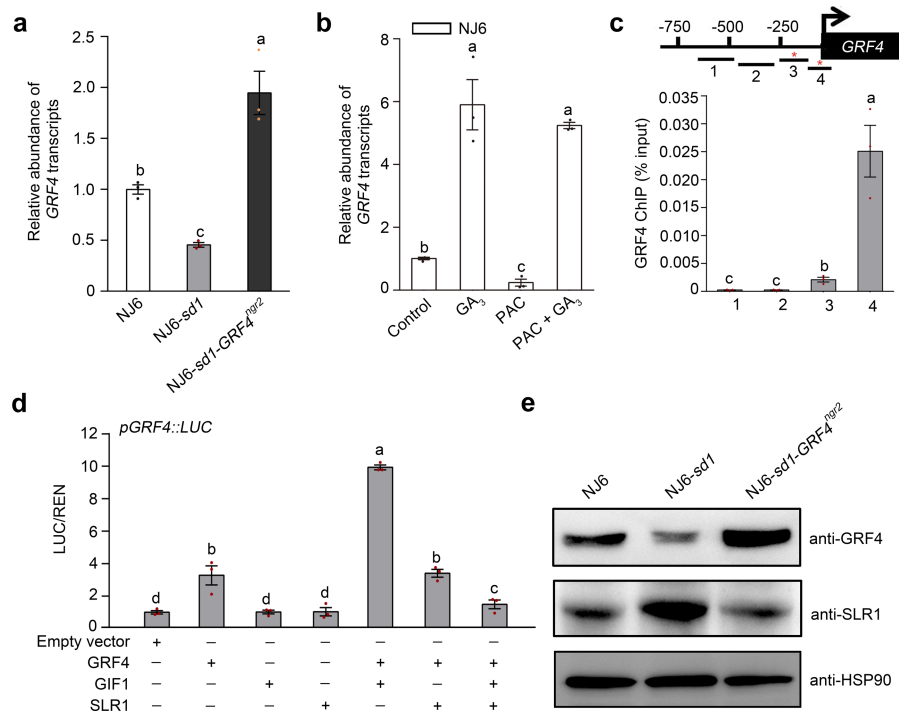
Extended Data Fig. 4 | GA promotes glutamine synthase and nitrate reductase activities. **a**, Glutamine synthase activities in roots of two-week-old rice plants treated with 100 μM GA (GA₃) and/or 2 μM PAC, genotypes as indicated. **b**, Glutamine synthase activities in shoots of plants

treated with GA and/or PAC, genotypes and treatments as indicated in **a**. **c**, Nitrate reductase activities in shoots of plants treated with GA and/or PAC, genotypes and treatments as indicated in **a**. **a–c**, Data are mean ± s.e.m. ($n = 3$); P values are from two-sided Student's t -tests.



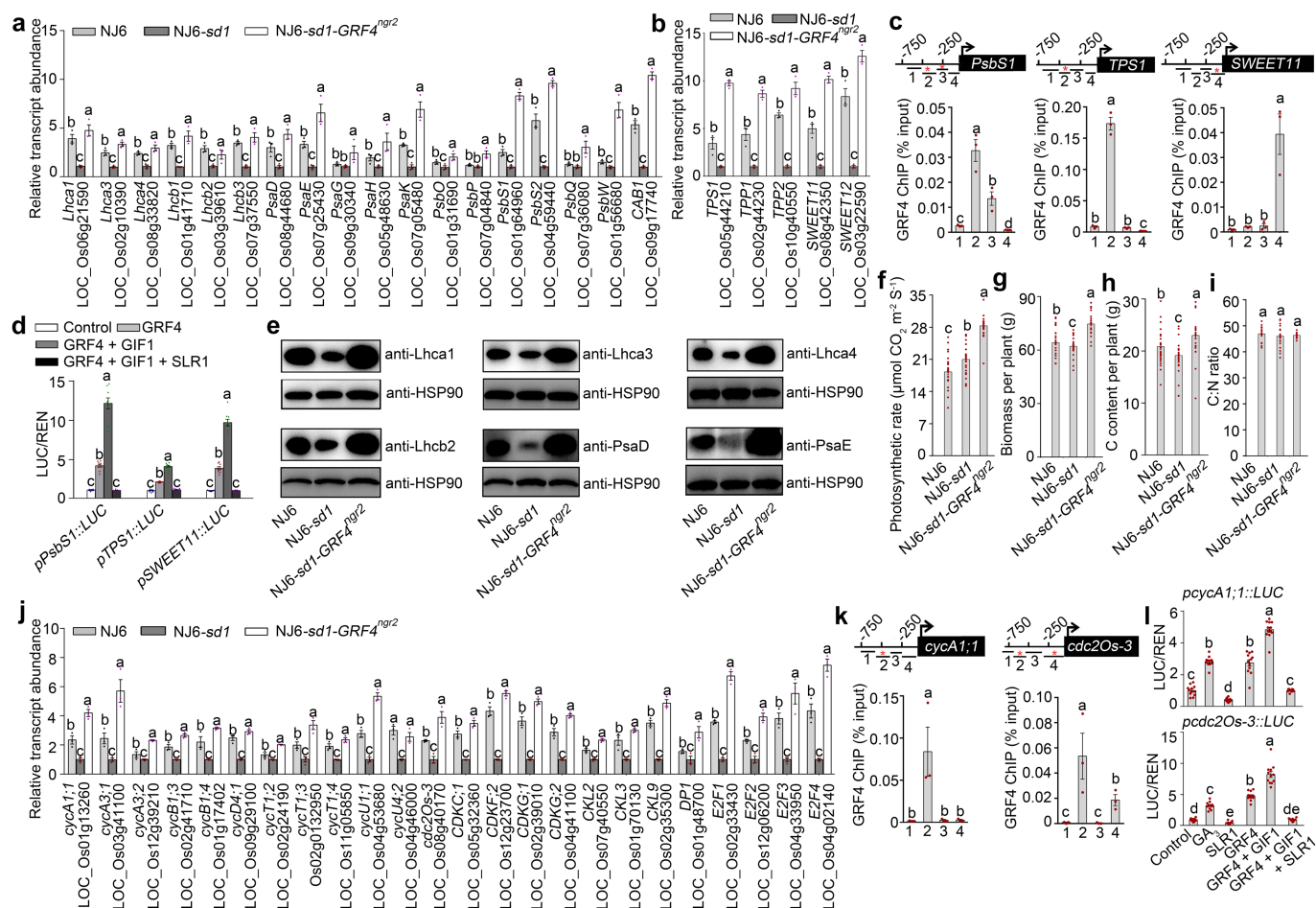
Extended Data Fig. 5 | BiFC visualization of SLR1–GIF1–GRF4 interactions. **a**, Details of constructs expressing GRF4 and variants deleted for specific domains. GRF4 contains the QLQ and WRC domains, positions as indicated. **b**, BiFC assays. Constructs expressing GRF4 or deletion variants (shown as in **a**) tagged with the N terminus of YFP were co-transformed into tobacco leaf epidermal cells, together with constructs expressing GIF1 or SLR1 tagged with the C terminus of YFP, respectively.

Scale bars, 60 μm . **c**, BiFC assays. Constructs expressing GRF1 or related GRFs and GIFs family proteins tagged with the N terminus of YFP were co-transformed into tobacco leaf epidermal cells together with a construct expressing SLR1 tagged with the C terminus of YFP. Scale bar, 60 μm . **b**, **c**, Images of BiFC assays are representative of three experiments performed independently with similar results.



Extended Data Fig. 6 | SLR1 inhibits GRF4-GIF1 self-promotion of *GRF4* mRNA and GRF4 protein abundance. **a**, *GRF4* mRNA abundance, plant genotypes as indicated. Abundances shown are relative to NJ6 (set to 1). **b**, The effects of GA and PAC on *GRF4* mRNA abundance in two-week-old NJ6 plants. Abundances shown are relative to the water treatment control (set to 1). **c**, ChIP-PCR *GRF4*-mediated enrichment (relative to input) of GCGG-containing *GRF4* promoter fragments (marked with asterisks). **d**, *GRF4*-activated promotion of transcription from the *GRF4*

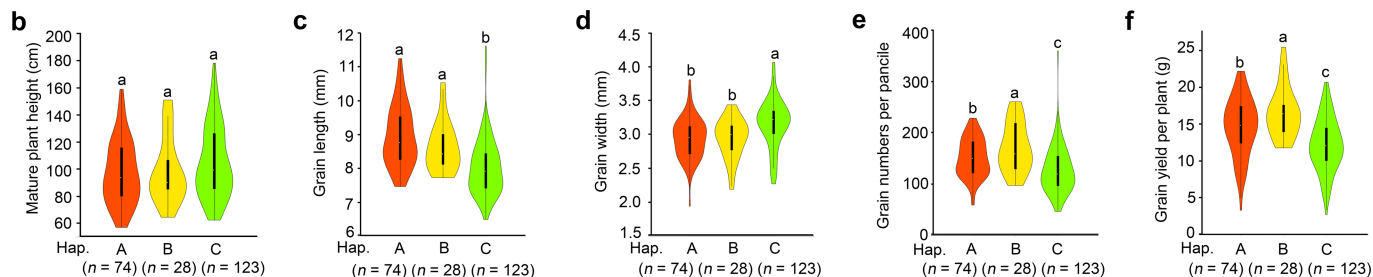
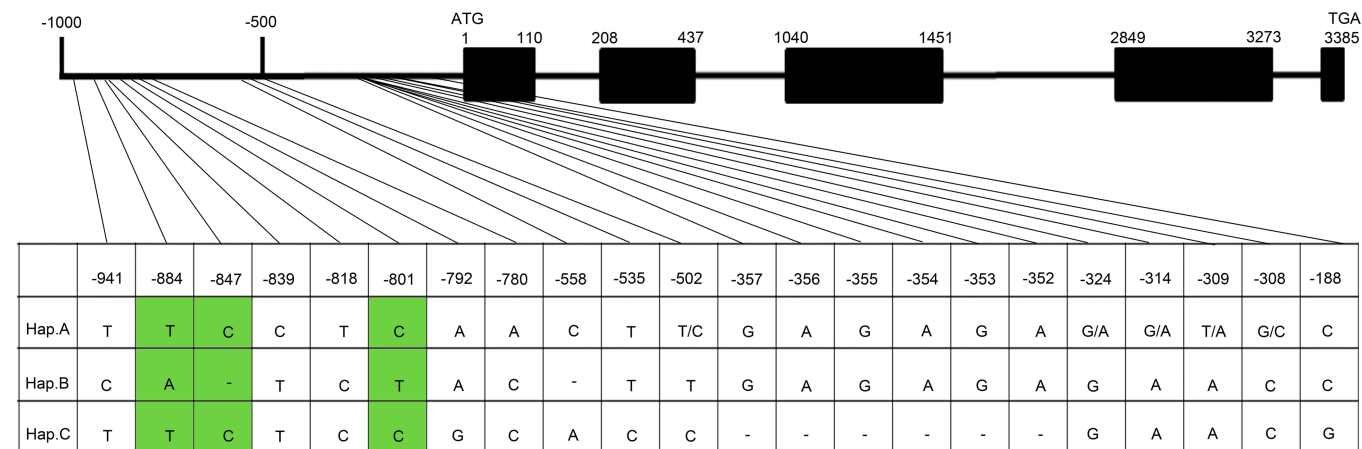
gene promoter-luciferase reporter construct is enhanced by GIF1 and inhibited by SLR1. Luciferase/renilla activity shown relative to the empty vector control (set to 1). **a-d**, Data are mean \pm s.e.m. ($n = 3$). Different letters denote significant differences ($P < 0.05$) from Duncan's multiple range tests. **e**, *GRF4* abundance (as detected by an anti-*GRF4* antibody), plant genotypes as indicated. HSP90 serves as loading control. Blots are representative of three experiments performed independently with similar results.



Extended Data Fig. 7 | The GRF4–SLR1 antagonism regulates carbon assimilation and plant growth. **a, b**, Relative shoot abundances of carbon-fixation gene mRNAs. Abundances of transcripts of genes regulating photosynthesis (**a**), sucrose metabolism and transport/phloem loading (**b**) in NJ6, NJ6-*sd1* and NJ6-*sd1-GRF4^{ngr2}* plants. Abundances in NJ6 and NJ6-*sd1-GRF4^{ngr2}* are expressed relative to NJ6-*sd1* (set to 1). **c**, ChIP–PCR assays. Diagrams depict the *PsbS1*, *TPS1* and *SWEET11* promoters and regions used for ChIP–PCR, and GCGG-containing promoter fragment (marked with asterisks) enrichment (relative to input). **a–c**, Data are mean \pm s.e.m. ($n = 3$). **d**, Transactivation assays. The luciferase/renilla activity obtained from a co-transfection with an empty effector construct and indicated reporter constructs was set to 1. Data are mean \pm s.e.m. ($n = 9$). **e**, Immunoblot detection of Lhca1, Lhca3, Lhca4, Lhcb2, PsaD and Psae using antibodies as shown in genotypes as indicated. HSP90 serves as

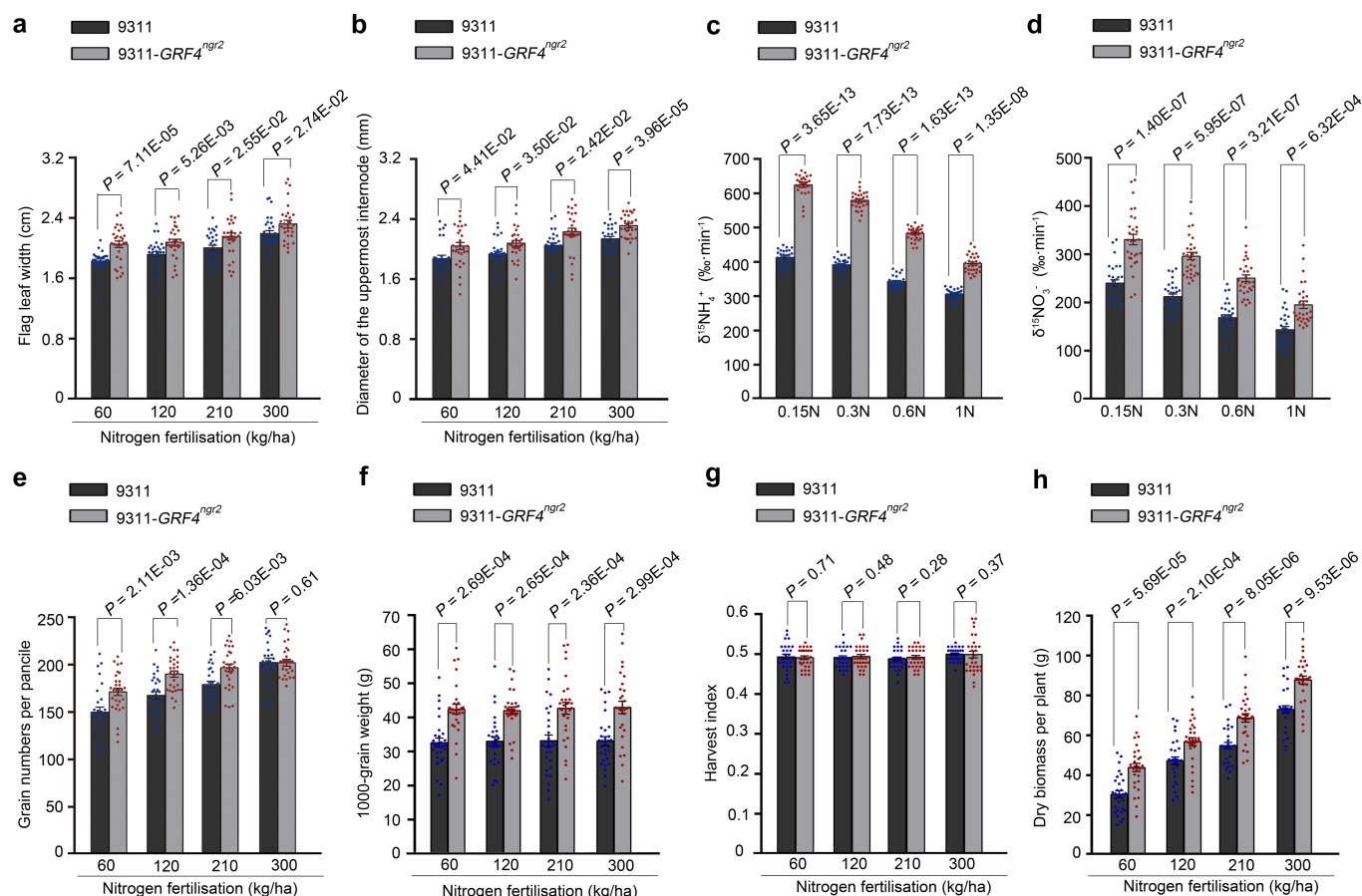
loading control. Blots are representative of three experiments performed independently with similar results. **f-i**, Comparisons of photosynthetic rates (**f**), biomass (**g**), carbon content (**h**) and C:N ratio (**i**) in NJ6, NJ6-*sd1* and NJ6-*sd1*-GRF4^{nr2} plants. Data are mean \pm s.e.m. ($n = 30$). **j**, Relative shoot abundances of mRNAs transcribed from cell-cycle regulatory genes in NJ6, NJ6-*sd1* and NJ6-*sd1*-GRF4^{nr2} plants. Transcription is relative to NJ6-*sd1* plants (set to 1). Data are mean \pm s.e.m. ($n = 3$). **k**, ChIP-PCR assays. Diagrams depict the *cycA1.1* and *cdc2Os-3* promoters and regions (CGCG-containing fragment marked with asterisks) used for ChIP-PCR. Data are mean \pm s.e.m. ($n = 3$). **l**, Transactivation assays from the *cycA1.1* and *cdc2Os-3* promoters. Data are mean \pm s.e.m. ($n = 12$). **a-d**, **f-l**, Different letters denote significant differences ($P < 0.05$) from Duncan's multiple range tests.

a LOC_Os02g47280



Extended Data Fig. 8 | Natural allelic variation at *GRF4* is associated with variation in plant and grain morphology and grain yield performance. **a**, DNA polymorphisms in the promoter region of *GRF4*. Green-shaded regions indicate the three unique SNP variations associated with phenotypic variation in NM73 and RD23. **b–f**, Box plots for plant height (**b**), grain length (**c**), grain width (**d**), the number of grains per panicle (**e**) and grain yield performance (**f**) of rice varieties carrying

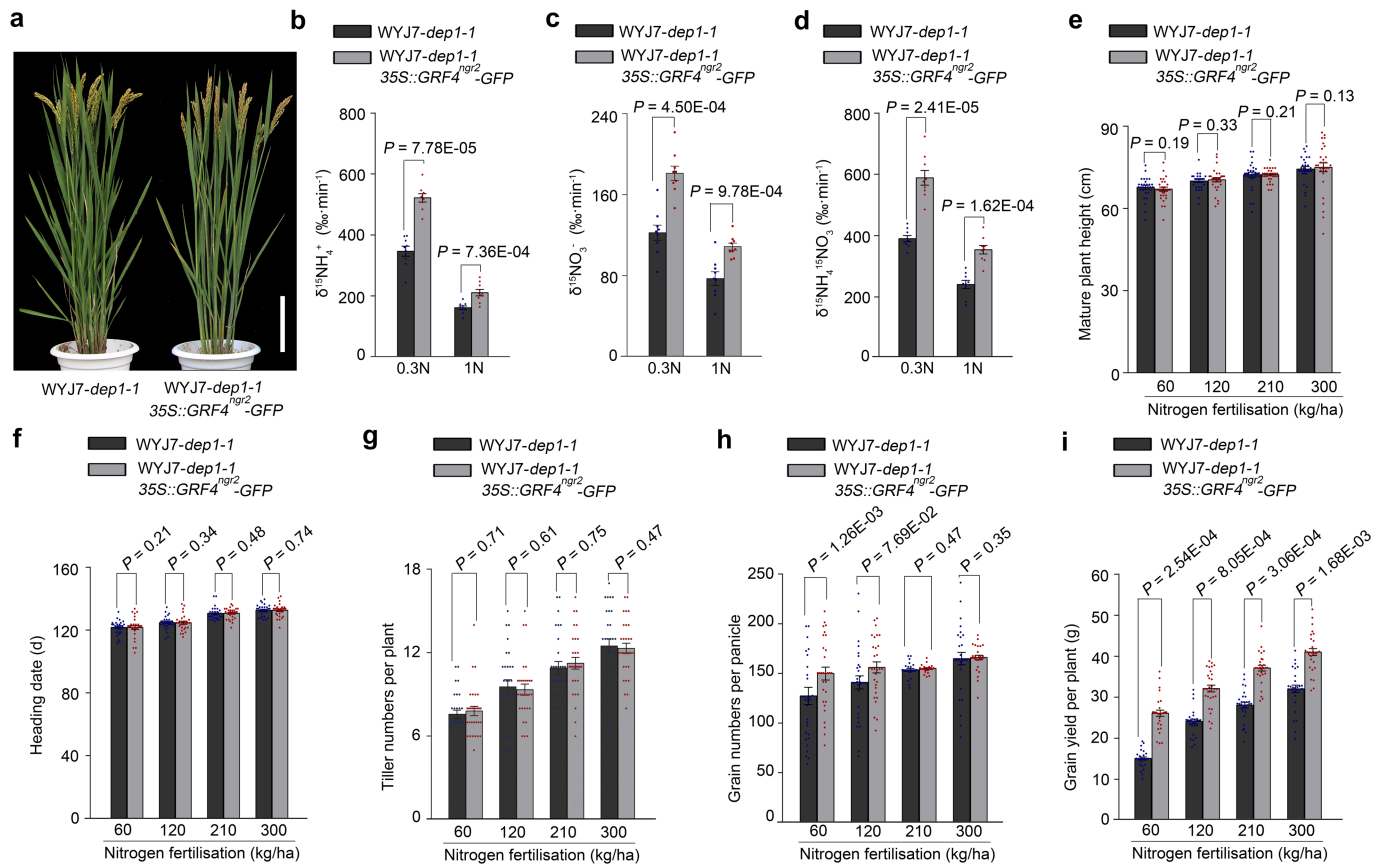
different *GRF4* promoter haplotypes (Hap. A, B or C). All data are from plants grown under normal paddy-field fertilization conditions²². Data are mean \pm s.e.m. (Hap. A, $n = 74$; Hap. B, $n = 28$; Hap. C, $n = 123$). The violin plot was constructed in R. **b–f**, Different letters indicate statistically significant differences between groups ($P < 0.05$) from a Tukey's honestly significant difference (HSD) test.



Extended Data Fig. 9 | Agronomic traits displayed by 9311 and 9311-GRF4^{ngr2} plants grown at varying nitrogen fertilization levels.

a, b, Flag leaf width (**a**) and culm width of the uppermost internode (**b**) at varying levels of nitrogen fertilization. **c, d**, $^{15}\text{NH}_4^+$ (**c**) and $^{15}\text{NO}_3^-$ (**d**) uptake rates of four-week-old plants grown with varying nitrogen supply

(0.15N, 0.1875 mM NH_4NO_3 ; 0.3N, 0.375 mM NH_4NO_3 ; 0.6N, 0.75 mM NH_4NO_3 ; 1N, 1.25 mM NH_4NO_3). **e–h**, The number of grains per panicle (**e**), 1,000-grain weight (**f**), harvest index (**g**) and dry biomass per plant (**h**) at varying levels of nitrogen fertilization. **a–h**, Data are mean \pm s.e.m. ($n = 30$); P values are from two-sided Student's t -tests.



Extended Data Fig. 10 | Growth, nitrogen uptake and grain yield performance of WYJ7-*dep1-1* and transgenic WYJ7-*dep1-1* plants carrying the 35S::GRF4^{ngr2}-GFP construct at varying levels of nitrogen fertilization. **a**, Mature plant heights. Scale bar, 15 cm. The picture is representative of three experiments performed independently with similar results. **b–d**, Root uptake rates for $^{15}\text{NH}_4^+$ (**b**), $^{15}\text{NO}_3^-$ (**c**), and $^{15}\text{NH}_4^+$ and $^{15}\text{NO}_3^-$ combined (**d**) of four-week-old rice plants grown in low nitrogen

(0.3N, 0.375 mM NH_4NO_3) and high nitrogen (1N, 1.25 mM NH_4NO_3) conditions. Data are mean \pm s.e.m. ($n = 9$). **e–i**, Mature plant height (**e**), heading date (**f**), the number of tillers per plant (**g**), the number of grains per panicle (**h**) and grain yield per plant (**i**) at varying levels of nitrogen fertilization. **e–i**, Data shown as mean \pm s.e.m. ($n = 30$). **b–i**, P values are from two-sided Student's t -tests.

Structure of paused transcription complex Pol II–DSIF–NELF

Seychelle M. Vos^{1,4}, Lucas Farnung^{1,4}, Henning Urlaub^{2,3} & Patrick Cramer^{1*}

Metazoan gene regulation often involves the pausing of RNA polymerase II (Pol II) in the promoter–proximal region. Paused Pol II is stabilized by the protein complexes DRB sensitivity–inducing factor (DSIF) and negative elongation factor (NELF). Here we report the cryo–electron microscopy structure of a paused transcription elongation complex containing *Sus scrofa* Pol II and *Homo sapiens* DSIF and NELF at 3.2 Å resolution. The structure reveals a tilted DNA–RNA hybrid that impairs binding of the nucleoside triphosphate substrate. NELF binds the polymerase funnel, bridges two mobile polymerase modules, and contacts the trigger loop, thereby restraining Pol II mobility that is required for pause release. NELF prevents binding of the anti–pausing transcription elongation factor IIS (TFIIS). Additionally, NELF possesses two flexible ‘tentacles’ that can contact DSIF and exiting RNA. These results define the paused state of Pol II and provide the molecular basis for understanding the function of NELF during promoter–proximal gene regulation.

Pol II transcribes eukaryotic protein-coding genes and is controlled at several levels. When Pol II begins to elongate the pre-mRNA chain, its activity is regulated by elongation factors with positive and negative roles¹. During the elongation of pre-mRNA, Pol II can be blocked^{2,3} or paused in the promoter–proximal region^{4,5}. Paused Pol II is stabilized by two factors: the 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) sensitivity-inducing factor (DSIF), composed of subunits SPT4 and SPT5, and the negative elongation factor (NELF), composed of the four subunits NELF-A, -B, -C (or isoform NELF-D that lacks the first nine NELF-C residues) and -E^{6–8}. Paused polymerase is released by the positive transcription elongation factor b (P-TEFb), which contains the kinase CDK9 and the predominant cyclin subunit CYCT1^{9,10}. P-TEFb phosphorylates Pol II, DSIF and NELF^{11–13}. DSIF and its homologues are conserved from bacteria to human, whereas NELF is generally conserved among metazoa⁸. Genes regulated by Pol II pausing often function during the development of an organism or during environmental responses^{14–17}. Pol II pausing is also used by viruses such as the human immunodeficiency virus (HIV)-1 to recruit viral factors such as Tat and to promote transcription elongation through P-TEFb^{3,10}.

The structural basis for RNA chain elongation by Pol II has been well studied¹⁸, but the factor-dependent mechanisms that regulate Pol II elongation are poorly understood. Our laboratory recently reported the structure of the mammalian Pol II elongation complex (EC) bound to DSIF¹⁹ and others reported a similar structure for yeast²⁰. DSIF was observed to form DNA and RNA clamps around the Pol II upstream DNA and the exiting RNA, respectively¹⁹. We also reported the crystal structure of a dimeric NELF subcomplex comprising the N-terminal region of NELF-A and the middle and C-terminal regions of NELF-C (NELF-A–NELF-C dimer)²¹. Other regions of NELF are structurally uncharacterized, except for the small RNA recognition motif domain of NELF-E²². It is not known how NELF binds the Pol II–DSIF EC and how it stabilizes pausing.

Here we provide the cryo-electron microscopy (cryo-EM) structure of a complex in which paused Pol II is bound to DSIF and NELF (termed the paused elongation complex, or PEC) at a nominal resolution of 3.2 Å. The structure was determined on a DNA–RNA scaffold that carries the sequence of a well-characterized promoter–proximal

pause site^{3,23} found in the HIV-1 provirus. Our structure indicates that NELF uses several mechanisms to interfere with nucleotide addition and polymerase progression. Complementary biochemical data and comparisons with published results provide the molecular basis for NELF-dependent promoter–proximal pausing.

Structure of the paused elongation complex

We purified endogenous porcine (*S. scrofa*) Pol II, which is nearly identical to human Pol II (Extended Data Table 1), and prepared recombinant human DSIF and NELF by co-expression of their subunits in bacteria and insect cells, respectively (Extended Data Fig. 1a, Methods). We then investigated transcriptional pausing in vitro with the use of a DNA template bearing a pause sequence and a short complementary RNA transcript (hereafter denoted the pause assay scaffold, Extended Data Fig. 1b, Methods). The sequence is well-studied in the bacterial transcription system and is known to cause pausing of mammalian Pol II²⁴. ECs were assembled with the RNA 3′-end two nucleotides upstream of the consensus pause site. We found that Pol II briefly paused at the consensus pause site (+2 position) in the absence of factors, as observed previously²⁴ (Fig. 1a, Extended Data Fig. 1e). Incubation with DSIF slightly suppressed pausing, whereas NELF alone had no effect on pausing or RNA elongation. By contrast, addition of both DSIF and NELF resulted in strong pausing at the +2 position and impeded further RNA extension (Fig. 1b, Extended Data Fig. 1e). These results show that our recombinant factors can stabilize Pol II in the paused state.

To determine the cryo-EM structure of the PEC, it was assembled using a DNA–RNA scaffold that mimics the nucleic acid arrangement during human Pol II pausing on a strong HIV-1 promoter–proximal pause site²³ (hereafter denoted the HIV-1 pause scaffold, Extended Data Fig. 1c). This scaffold includes a hairpin in the exiting RNA known as the transactivation response element (TAR)³. A similar scaffold with a nearly identical template sequence recapitulates the effects of DSIF and NELF on Pol II pausing (hereafter denoted the HIV-1 transcription scaffold, Extended Data Fig. 2). The PEC was assembled from pure components, isolated by size-exclusion chromatography, and gently cross-linked with glutaraldehyde (Fig. 1c, Extended Data Fig. 1d).

¹Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Göttingen, Germany. ²Max Planck Institute for Biophysical Chemistry, Bioanalytical Mass Spectrometry, Göttingen, Germany. ³University Medical Center Göttingen, Institute of Clinical Chemistry, Bioanalytics Group, Göttingen, Germany. ⁴These authors contributed equally: Seychelle M. Vos, Lucas Farnung. *e-mail: patrick.cramer@mpibpc.mpg.de

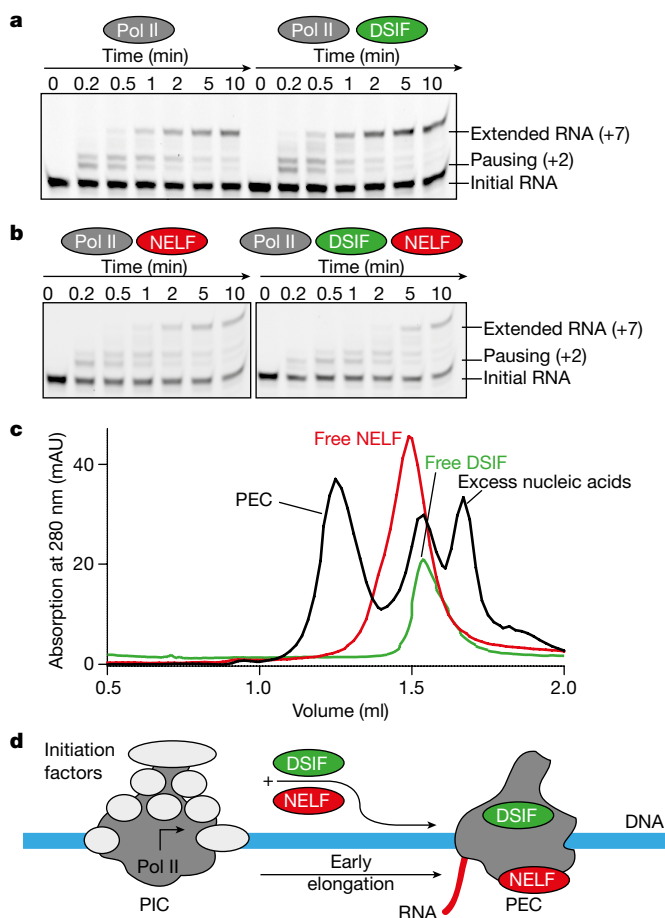


Fig. 1 | Formation of the paused Pol II-DSIF-NELF elongation complex. **a**, DSIF alone does not stabilize Pol II pausing. Fluorescence-monitored RNA extension (see Methods) on the pause assay scaffold (50 nM) with a 5' carboxyfluorescein (FAM)-labelled RNA, 75 nM Pol II (left) or 75 nM Pol II and 237 nM DSIF (right). Reactions were quenched at various times after the addition of GTP and CTP (10 μ M) and RNAs were separated on TBE urea gels. ECs were assembled two nucleotides before the consensus pause site (+2). The band above the +2 band arises from a backtracked species²⁴. The +2 site and extended RNA are marked. All experiments were performed at least three times. Quantification of gels in panels **a** and **b** can be found in Extended Data Fig. 1e. A fraction of the input RNA remains owing to inefficient EC formation (see Methods). **b**, DSIF and NELF are required for stable Pol II pausing. Experiments conducted as in **a** but in the presence of 237 nM NELF (left) or 237 nM DSIF and 237 nM NELF (right). All experiments were performed at least three times. **c**, Formation of a stable paused Pol II-DSIF-NELF EC (PEC) on a Superose 6 size-exclusion chromatography column. Curves show absorption at 280 nm milli absorption units (mAU) at specific elution volumes. All experiments were performed three times. **d**, Schematic of conversion of the Pol II pre-initiation complex (PIC) to a promoter-proximally paused Pol II-DSIF-NELF EC (PEC).

The cryo-EM structure was determined from 162,269 particles at a nominal resolution of 3.2 Å (Fig. 2, Extended Data Fig. 3, Supplementary Video 1). The Pol II core was resolved at 3.0 Å, whereas DSIF and NELF were resolved at local resolutions of around 3.5 Å to 8 Å (Extended Data Figs. 4, 5). We placed structures of Pol II, DSIF¹⁹ and the NELF-A-NELF-C dimer²¹ into our density maps and made minor adjustments (see Methods). We then extended the NELF-C model by tracing into continuous density and modelled the helical subunit NELF-B into density with well-defined secondary structure (Supplementary Table 1). Density for NELF-E was largely absent, except for the N-terminal helix α 1. The PEC structure was confirmed by chemical crosslinking (Methods, Extended Data Fig. 6, Supplementary Tables 2, 3) and shows good stereochemistry (Extended Data Table 2).

Tilting of the DNA-RNA hybrid

The PEC structure adopts the closed Pol II conformation that has been observed in the structures of other elongation complexes¹⁸ (Fig. 2). The Pol II structure is highly similar to that previously determined in the Pol II-DSIF EC¹⁹, with the exception of an approximately 10 Å movement in the flexible RPB4-RPB7 stalk and a slightly altered trajectory of upstream DNA. DSIF domains are arrayed around upstream DNA and exiting RNA, with minor movements of the KOW1 and NGN domains (Extended Data Fig. 7a). The downstream DNA and the DNA-RNA hybrid are very well defined, whereas only weak, uninterpretable density is observed for the TAR RNA hairpin.

The DNA-RNA hybrid in the Pol II active site adopts a tilted conformation (Fig. 3, Supplementary Video 2). This unusual hybrid conformation has been previously observed in Pol II complexes containing backtracked RNA²⁵ or short DNA-RNA hybrids²⁶. The tilted conformation occurs in an off-line state that is not part of the productive nucleotide addition cycle. It may be adopted during transcription if the post-translocated state is unstable and the DNA, but not the RNA, slides backwards to the pre-translocation position while maintaining DNA-RNA base pairing. As a consequence, the RNA adopts a post-translocated position, whereas the DNA appears pre-translocated. The tilted state readily explains polymerase pausing, because there is no free DNA template base in the active site that could bind the nucleotide triphosphate (NTP) by canonical base pairing (Fig. 3b). A tilted hybrid was also recently observed in paused bacterial ECs^{27,28}, and therefore probably underlies the fundamental paused state of cellular multi-subunit RNA polymerases.

NELF adopts a three-lobed structure

NELF forms a three-lobed structure that binds to Pol II on the face opposite the cleft (Figs. 2, 4). One lobe is formed by the NELF-A-NELF-C dimer (hereafter the NELF-A-NELF-C lobe), which adopts a more open conformation relative to the free structure (Extended Data Fig. 7b). A second lobe (the NELF-B-NELF-C lobe) is formed by the N-terminal regions of NELF-B (α 1- α 6) and NELF-C (α N1- α N9). This lobe comprises two pairs of helices formed by NELF-C (α N4- α N5) and NELF-B (α 3- α 4) that resemble open pairs of scissors (Fig. 4) and a four-helix bundle encompassing NELF-C α N2- α N3 and NELF-B α 1- α 2. A third lobe (the NELF-B-NELF-E lobe) is formed by the NELF-B 'staircase' and HEAT (huntingtin, elongation factor 3, protein phosphatase 2A and TOR1) domains and by NELF-E helix α 1, which lies between these NELF-B domains. The NELF-B staircase shares modest structural similarity with yeast exportin 1 (residues 388-638), and the NELF-B HEAT domain comprises four canonical HEAT repeats (see Methods).

The three lobes of NELF are well conserved²¹ (Supplementary Tables 4, 5) from *Dictyostelium discoideum* to human, which suggests that our structure is a good model for all NELF homologues. Most of NELF-B is nearly identical among vertebrates, whereas the HEAT domain shows some sequence divergence. The mobile C-terminal regions of NELF-A and NELF-E are less conserved²¹ (Supplementary Tables 4, 5). Taken together, these results show that four NELF subunits interact extensively to form a compact three-lobed structure, which is predicted to be highly stable and conserved.

NELF restrains Pol II mobility

The NELF-A-NELF-C lobe docks on the rim of the Pol II funnel that leads to the pore and the active site²⁹ (Figs. 2, 5). The funnel and pore together are known as the 'secondary channel' in bacterial polymerase³⁰ and are suggested to provide an entry route for NTP substrates. On the rim of the funnel, NELF-A contacts RPB8, and NELF-C interacts with the RPB1 funnel helices (α 20, α 21) and the RPB1 cleft domain (α 38) near the foot (Extended Data Fig. 8a, b). The NELF-Pol II interface is highly charged (Extended Data Table 3). The NELF-interacting polymerase regions reside in two different mobile modules of Pol II²⁹. RPB8 and the RPB1 funnel domain reside in the core module, whereas the RPB1 cleft and foot domains reside in the shelf module. NELF binding

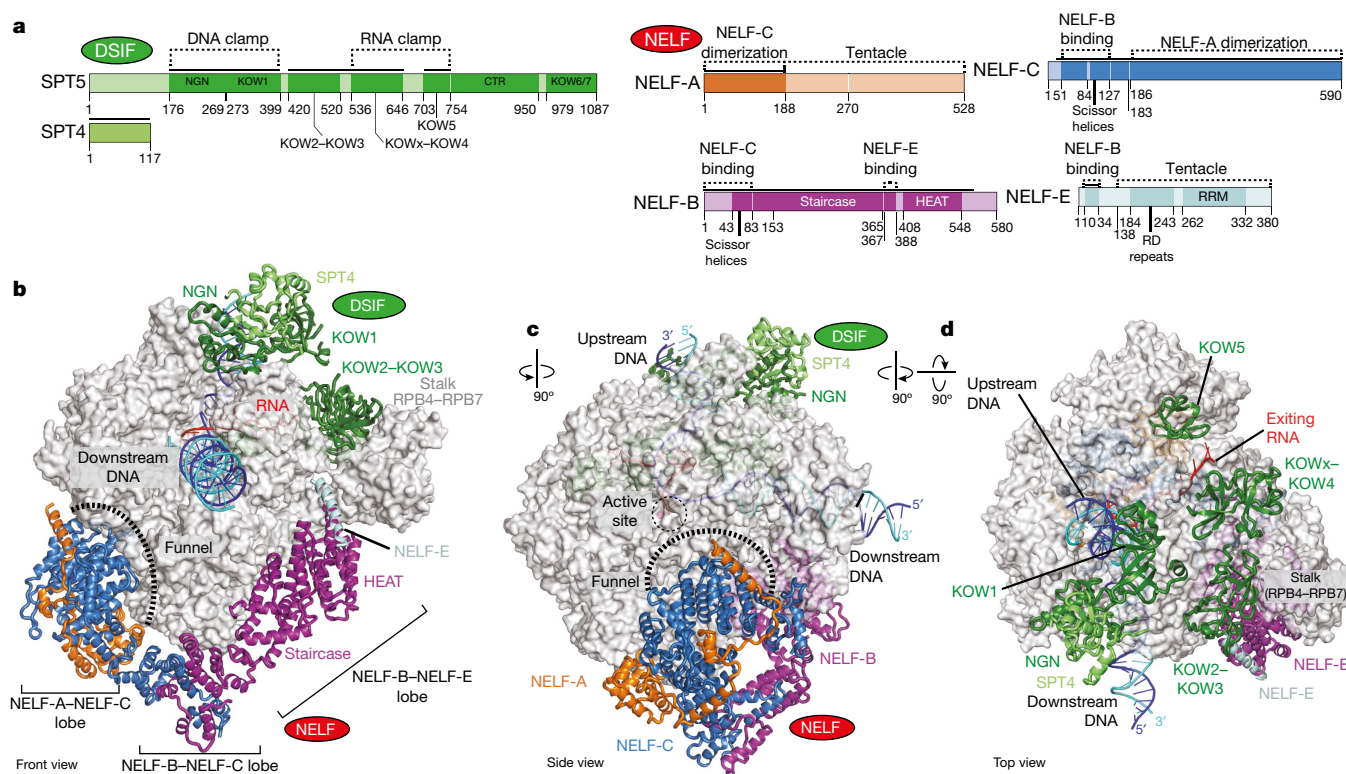


Fig. 2 | Cryo-EM structure of the PEC. **a**, Domain architecture of DSIF and NELF subunits. The colour code is used throughout all figures. Solid black lines indicate modelled regions. **b–d**, Cartoon model of the PEC viewed from the Pol II front (**b**), side (**c**) and top (**d**). Pol II is shown as

a silver surface, DSIF and NELF are shown as ribbon models. The active site metal ion A is depicted as a magenta sphere, DNA template and non-template strands are in blue and cyan, respectively, and RNA is shown in red.

to the funnel does not change the positions of the core and shelf modules compared to the Pol II–DSIF EC structure¹⁹. NELF binding, however, is predicted to restrain the relative movement of these polymerase modules, which occurs during Pol II reactivation from an arrested state^{25,31} (Extended Data Fig. 9a).

The NELF-A–NELF-C lobe additionally contacts the open Pol II trigger loop (Fig. 5a). In particular, a loop connecting NELF-C helices $\alpha 17$ and $\alpha 18$ lies in close proximity to the tip of the open trigger loop (Extended Data Figs. 5b, 8c). The trigger loop is a highly conserved element of the polymerase active site that generally adopts an open, mobile conformation, but folds and closes over the incoming NTP for catalytic RNA chain extension^{32,33}. Although the observed contact is not extensive, it could impair trigger loop closure and catalysis.

The NELF-B–NELF-C lobe does not associate with Pol II, whereas the NELF-B–NELF-E lobe forms a few contacts. The NELF-B staircase domain contacts Pol II subunits RPB5 and RPB6. NELF-B helix $\alpha 13$

interacts with a negatively charged loop in the RPB5 jaw domain (loop $\beta 1$ – $\beta 2$). NELF-B helices $\alpha 23$ – $\alpha 26$ in the HEAT domain reside near the RPB6 N-terminal region. The NELF-E helix $\alpha 1$ lies adjacent to the base of the outer Pol II clamp. Generally, Pol II regions contacted by NELF do not appear to be conserved in Pol I and Pol III, which indicates that NELF function is specific to Pol II (Extended Data Fig. 8d). Taken together, these results show that NELF interacts with Pol II via two of its three lobes. The NELF-A–NELF-C lobe specifically forms contacts with the Pol II funnel and trigger loop that restrain Pol II mobility, and is likely to stabilize the tilted conformation of the hybrid and therefore the paused state.

Two NELF ‘tentacles’ reach DSIF and RNA

Two mobile regions extend from the NELF body and contact DSIF (Fig. 6), potentially explaining why NELF function requires DSIF^{7,8}. We refer to these two extensions as ‘tentacles’: the NELF-A tentacle (residues 189–528) and the NELF-E tentacle (residues 139–363). Weak density (not shown) and crosslinking data indicate that the NELF-A tentacle extends from the NELF-A–NELF-C lobe along the RPB2 protrusion to reach the DSIF DNA clamp, in particular SPT4 and the SPT5 NGN domain (Fig. 6a, Extended Data Fig. 7c). The NELF-A tentacle is important for Pol II binding⁸ and overlaps with a region of TFIIF that binds this surface of Pol II³⁴. Additionally, crosslinking data suggest that the NELF-E tentacle extends from helix $\alpha 1$ across the DSIF KOW2–KOW3 and KOWx–KOW4 domains towards exiting RNA (Fig. 6b, Extended Data Fig. 7d).

To test whether the NELF tentacles are involved in pause stabilization, we prepared truncated NELF variants and performed RNA extension assays on the pause assay scaffold and the HIV-1 transcription scaffold (see Methods). A variant lacking the NELF-A tentacle could not stabilize the pause, whereas a variant lacking the NELF-E tentacle was functional (Fig. 6c, Extended Data Figs. 1f, 2d, e). This shows that the NELF-A tentacle is required for NELF function in pause stabilization, consistent with published data⁸, whereas the NELF-E tentacle is not.

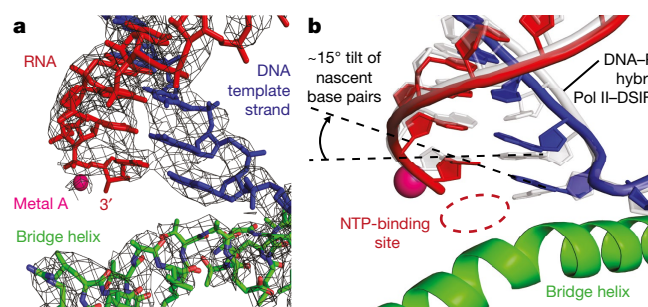


Fig. 3 | Tilted DNA–RNA hybrid. **a**, Cryo-EM density (grey mesh) for the DNA–RNA hybrid and bridge helix in the Pol II active site. **b**, Comparison of the tilted DNA–RNA hybrid in the PEC structure (blue and red) with the post-translocated hybrid in the previously solved structure of the Pol II–DSIF EC¹⁹ (grey).

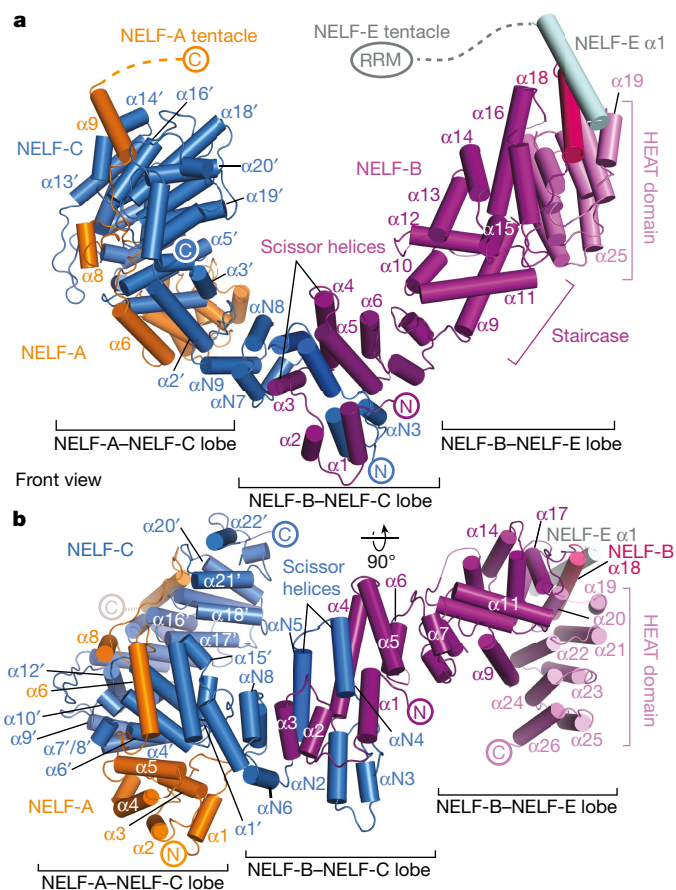


Fig. 4 | NELF structure. **a**, Three-lobed NELF structure adopted in the PEC. Helices are shown as cylinders. NELF domains and elements are indicated. The view corresponds to the front view in Fig. 2b. **b**, The structure shown in **a**, rotated by 90° around a horizontal axis.

The NELF-E tentacle encompasses the RNA recognition motif domain, which is not required for Pol II association³⁵ but can bind RNA hairpins³⁶. RNA hairpin structures are enriched at strong pause sites³⁷, but RNA binding by NELF is not required for pausing³⁸. Therefore, the NELF-E tentacle is not required for pausing but may bind nascent RNA to help to recruit NELF to pause sites.

NELF impairs TFIIS binding

Transcriptional pausing involves polymerase stalling but can additionally involve the backtracking of polymerase on DNA and RNA³⁹. Rescue of backtracked Pol II requires TFIIS, which stimulates cleavage of the nascent RNA 3'-end⁴⁰. Regions of the genome that are prone to backtracking are also susceptible to promoter-proximal pausing and require TFIIS for pause release^{41,42}. As DSIF and NELF were reported to inhibit TFIIS-stimulated RNA cleavage⁴³, we compared our PEC structure with previous Pol II–TFIIS structures^{31,44}. Superposition shows that the location of the NELF-A–NELF-C lobe is incompatible with TFIIS binding to the Pol II funnel (Fig. 5b). In particular, NELF is predicted to impair entry of the TFIIS interdomain linker between the polymerase core and shelf modules. We therefore tested whether TFIIS could bind the PEC in vitro. TFIIS was unable to form a complex with the PEC, although TFIIS readily bound the Pol II–DSIF EC (Extended Data Fig. 9b, c). These data show that binding of NELF and TFIIS to the funnel is mutually exclusive and suggest that NELF impairs TFIIS-mediated reactivation of Pol II.

Discussion

Here we formed a paused Pol II–DSIF–NELF elongation complex and resolved its structure. The PEC structure contains a tilted DNA–RNA hybrid that is incompatible with binding of the NTP substrate. It also

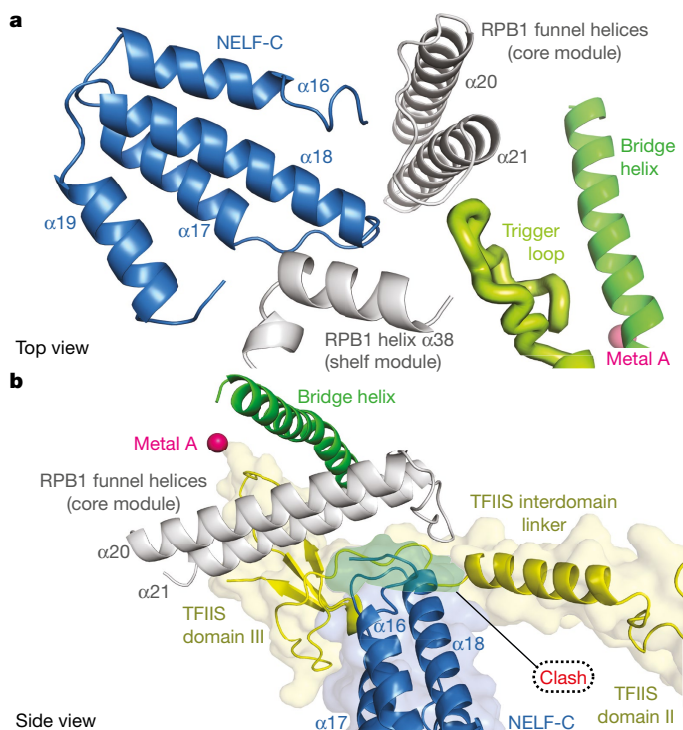


Fig. 5 | NELF restricts Pol II mobility. **a**, The NELF-A–NELF-C dimer bridges the Pol II core and shelf modules. The loop connecting NELF-C helices 17 and 18 contacts the open trigger loop (light green). **b**, NELF sterically impairs TFIIS binding. A human Pol II–TFIIS structure (PDB ID: 5IYC)⁴⁴ was superimposed onto the PEC structure by matching the Pol II core modules. TFIIS is shown in yellow, and the clashing region is shown in green.

revealed that NELF comprises three structured lobes and two flexible tentacles that approach DSIF and exiting RNA. Our results suggest five possible mechanisms that NELF may use to stabilize pausing in an allosteric manner; that is, without reaching the Pol II active site. First, binding of NELF along the Pol II funnel restricts movements of the two major polymerase modules, core and shelf, which may stabilize the tilted state of the hybrid. Second, NELF restricts the funnel and may therefore interfere with NTP diffusion into the funnel and reduce substrate delivery to the active site. Third, NELF contacts the open trigger loop, which may hinder its closure and nucleotide addition. Fourth, NELF interferes with TFIIS binding, therefore impeding reactivation of Pol II by TFIIS, which involves movement of the shelf module with respect to the core module^{25,31}. Finally, NELF could sterically or allosterically block binding of other positive elongation factors.

Together with published results, our data suggest that the nature of the paused state is likely to be the same for all multi-subunit cellular RNA polymerases. Nucleic acid sequences that can lead to pausing are conserved from bacteria to human^{45,46}, and a bacterial pause sequence can induce pausing of mammalian Pol II²⁴. It was recently reported that paused bacterial polymerase complexes contain a tilted DNA–RNA hybrid^{27,28}, as shown here for the mammalian PEC structure. These observations suggest that the paused state is conserved. Pausing by bacterial and eukaryotic polymerases is, however, differentially influenced by flanking DNA sequences^{45,47,48}. Some DNA sequences give rise to hairpins in nascent RNA that can be bound directly by bacterial RNA polymerase within the RNA exit tunnel^{27,28}. By contrast, RNA hairpin binding by the Pol II exit tunnel is probably prevented by DSIF¹⁹. An RNA hairpin may however form on the Pol II surface near the RPB4–RPB7 stalk, where it could be bound by the NELF-E tentacle.

Promoter-proximal pausing follows transcription initiation (Fig. 1d). Our results show that NELF can stabilize pausing only after initiation factors have been released. Modelling shows that the binding locations

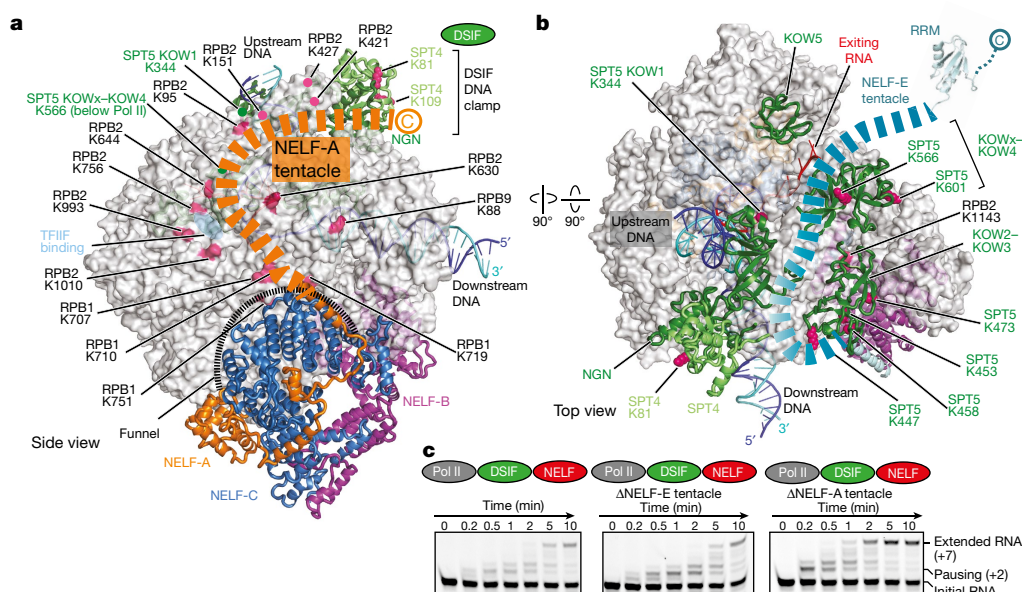


Fig. 6 | NELF tentacles reach DSIF and RNA. a, The NELF-A tentacle. Residues 189–528 of NELF-A form a flexible tentacle that binds Pol II and DSIF. Lysine crosslinking sites are marked. **b**, The NELF-E tentacle. Residues 139–363 of NELF-E form a flexible tentacle that extends over DSIF, close to the exiting RNA. **c**, The NELF-A tentacle, but not the

NELF-E tentacle, is required for pause stabilization. RNA extension assays were performed as in Fig. 1. All experiments were performed at least three times. Quantification of gels can be found in Extended Data Fig. 1f. RRM, RNA recognition motif.

of DSIF and NELF on Pol II are shared with some initiation factors⁴⁴. DSIF binding is incompatible with TFIIB and TFIIE, whereas NELF-A is probably mutually exclusive with TFIIF, which explains previous biochemical results⁴⁹. After dissociation of the initiation factors, the association of DSIF and NELF may not only stabilize the paused state but also prevent reassociation of initiation factors. Taken together, our results establish how NELF associates with a paused Pol II–DSIF EC and how bound NELF can stabilize pausing. Understanding these NELF interactions also helps to define how NELF can be dissociated when paused Pol II is activated for efficient elongation. In the accompanying paper⁵⁰, we describe how NELF can be released and how an activated Pol II EC is formed.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0442-2>.

Received: 19 March 2018; Accepted: 17 July 2018;

Published online 22 August 2018.

- Shilatfard, A., Conaway, R. C. & Conaway, J. W. The RNA polymerase II elongation complex. *Annu. Rev. Biochem.* **72**, 693–715 (2003).
- Bentley, D. L. & Groudine, M. A block to elongation is largely responsible for decreased transcription of *c-myc* in differentiated HL60 cells. *Nature* **321**, 702–706 (1986).
- Kao, S.-Y., Calman, A. F., Luciw, P. A. & Peterlin, B. M. Anti-termination of transcription within the long terminal repeat of HIV-1 by *tat* gene product. *Nature* **330**, 489–493 (1987).
- Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced *hsp70* gene in *Drosophila melanogaster* cells. *Mol. Cell. Biol.* **6**, 3984–3989 (1986).
- Strobl, L. J. & Eick, D. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of *c-myc* in vivo. *EMBO J.* **11**, 3307–3314 (1992).
- Wada, T. et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* **12**, 343–356 (1998).
- Yamaguchi, Y. et al. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41–51 (1999).
- Narita, T. et al. Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Mol. Cell. Biol.* **23**, 1863–1873 (2003).
- Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270**, 12335–12338 (1995).
- Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451–462 (1998).
- Fujinaga, K. et al. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol. Cell. Biol.* **24**, 787–795 (2004).
- Yamada, T. et al. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol. Cell* **21**, 227–237 (2006).
- Marshall, N. F., Peng, J., Xie, Z. & Price, D. H. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J. Biol. Chem.* **271**, 27176–27183 (1996).
- Rahl, P. B. et al. c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010).
- Williams, L. H. et al. Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol. Cell* **58**, 311–322 (2015).
- Adelman, K. et al. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc. Natl Acad. Sci. USA* **106**, 18207–18212 (2009).
- Wu, C.-H. et al. NELF and DSIF cause promoter proximal pausing on the *hsp70* promoter in *Drosophila*. *Genes Dev.* **17**, 1402–1414 (2003).
- Martinez-Rucobo, F. W. & Cramer, P. Structural basis of transcription elongation. *Biochim. Biophys. Acta* **1829**, 9–19 (2013).
- Bernecky, C., Plitzko, J. M. & Cramer, P. Structure of a transcribing RNA polymerase II–DSIF complex reveals a multidentate DNA–RNA clamp. *Nat. Struct. Mol. Biol.* **24**, 809–815 (2017).
- Ehara, H. et al. Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science* **357**, 921–924 (2017).
- Vos, S. M. et al. Architecture and RNA binding of the human negative elongation factor. *eLife* **5**, e14981 (2016).
- Rao, J. N. et al. Structural studies on the RNA-recognition motif of NELF E, a cellular negative transcription elongation factor involved in the regulation of HIV transcription. *Biochem. J.* **400**, 449–456 (2006).
- Palangat, M., Meier, T. I., Keene, R. G. & Landick, R. Transcriptional pausing at +62 of the HIV-1 nascent RNA modulates formation of the TAR RNA structure. *Mol. Cell* **1**, 1033–1042 (1998).
- Larson, M. H. et al. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042–1047 (2014).
- Cheung, A. C. M. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249–253 (2011).
- Cheung, A. C. M., Sainsbury, S. & Cramer, P. Structural basis of initial RNA polymerase II transcription. *EMBO J.* **30**, 4755–4763 (2011).
- Kang, J. Y. et al. RNA polymerase accommodates a pause RNA hairpin by global conformational rearrangements that prolong pausing. *Mol. Cell* **69**, 802–815.e1 (2018).
- Guo, X. et al. Structural basis for NusA stabilized transcriptional pausing. *Mol. Cell* **69**, 816–827.e4 (2018).

29. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science* **292**, 1863–1876 (2001).
30. Zhang, G. et al. Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98**, 811–824 (1999).
31. Kettenberger, H., Armache, K.-J. & Cramer, P. Architecture of the RNA polymerase II–TFIIS complex and implications for mRNA cleavage. *Cell* **114**, 347–357 (2003).
32. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941–954 (2006).
33. Vassylyev, D. G., Vassylyeva, M. N., Perederina, A., Tahir, T. H. & Artsimovitch, I. Structural basis for transcription elongation by bacterial RNA polymerase. *Nature* **448**, 157–162 (2007).
34. Plaschka, C. et al. Transcription initiation complex structures elucidate DNA opening. *Nature* **533**, 353–358 (2016).
35. Yamaguchi, Y., Inukai, N., Narita, T., Wada, T. & Handa, H. Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA. *Mol. Cell. Biol.* **22**, 2918–2927 (2002).
36. Pagano, J. M. et al. Defining NELF-E RNA binding in HIV-1 and promoter-proximal pause regions. *PLoS Genet.* **10**, e1004090 (2014).
37. Gressel, S. et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife* **6**, e29736 (2017).
38. Missra, A. & Gilmour, D. S. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the *Drosophila* RNA polymerase II transcription elongation complex. *Proc. Natl Acad. Sci. USA* **107**, 11301–11306 (2010).
39. Landick, R. The regulatory roles and mechanism of transcriptional pausing. *Biochem. Soc. Trans.* **34**, 1062–1066 (2006).
40. Reines, D., Ghanouni, P., Li, Q. Q. & Mote, J., Jr. The RNA polymerase II elongation complex. Factor-dependent transcription elongation involves nascent RNA cleavage. *J. Biol. Chem.* **267**, 15516–15522 (1992).
41. Adelman, K. et al. Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol. Cell* **17**, 103–112 (2005).
42. Nechaev, S. et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338 (2010).
43. Palangat, M., Renner, D. B., Price, D. H. & Landick, R. A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS. *Proc. Natl Acad. Sci. USA* **102**, 15036–15041 (2005).
44. He, Y. et al. Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359–365 (2016).
45. Bochkareva, A., Yuzenkova, Y., Tadigotla, V. R. & Zenkin, N. Factor-independent transcription pausing caused by recognition of the RNA–DNA hybrid sequence. *EMBO J.* **31**, 630–639 (2012).
46. Imashimizu, M. et al. Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.* **16**, 98 (2015).
47. Kireeva, M. L. & Kashlev, M. Mechanism of sequence-specific pausing of bacterial RNA polymerase. *Proc. Natl Acad. Sci. USA* **106**, 8900–8905 (2009).
48. Palangat, M., Hittinger, C. T. & Landick, R. Downstream DNA selectively affects a paused conformation of human RNA polymerase II. *J. Mol. Biol.* **341**, 429–442 (2004).
49. Cheng, B. & Price, D. H. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J. Biol. Chem.* **282**, 21901–21912 (2007).
50. Vos, S. M. et al. Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* <https://doi.org/10.1038/s41586-018-0440-4> (2018).

Acknowledgements We thank E. Wolf for pig thymus, F. Fischer and U. Neef for maintaining insect cell stocks, C. Oberthür and G. Kocic for assistance with protein purification, A. Linden and C.-T. Lee for help with crosslinking mass spectrometry, C. Bernecky for discussions and for sharing the DSIF plasmid before publication, and D. Tegenov and C. Wigge for electron microscopy support. S.M.V. was supported by an EMBO Long-Term Fellowship (ALTF 745-2014). H.U. was supported by the Deutsche Forschungsgemeinschaft (DFG SFB860). P.C. was supported by the Advanced Grant TRANSREGULON (grant agreement 693023) of the European Research Council, and the Volkswagen Foundation.

Reviewer information *Nature* thanks K. Adelman, S. Darst and R. Landick for their contribution to the peer review of this work.

Author contributions S.M.V. designed and conducted all experiments, unless stated otherwise. L.F. collected and processed electron microscopy data. L.F. and S.M.V. built the model. H.U. performed mass spectrometry. P.C. supervised the project. S.M.V., L.F. and P.C. wrote the manuscript with input from H.U.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0442-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0442-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and protein expression. DSIF was expressed in bacteria as described¹⁹. *Escherichia coli* expressing DSIF were collected by centrifugation, resuspended in Lysis 500 buffer (500 mM NaCl, 50 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 50 mM imidazole pH 8.0, 1 mM dithiothreitol (DTT), 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF, and 0.33 mg ml^{-1} benzamidine), flash-frozen in liquid nitrogen, and stored at -80°C until purification.

NELF was cloned and expressed as previously described²¹. The construct contains NELF-D, which is identical to NELF-C but lacks the first nine amino acid residues. For simplicity, NELF-C numbering is used throughout the manuscript, unless otherwise stated. The NELF tentacle variants (NELF-A 1–188 and NELF-E 1–138) were cloned by round-the-horn site-directed mutagenesis and incorporated into a single baculovirus expression vector containing the remaining three subunits by ligation independent cloning⁵¹. Sf9 (ThermoFisher), Sf21 (Expression Systems), and Hi5 (Expression Systems) cell lines were not tested for mycoplasma contamination and were not authenticated in-house. Hi5 cells expressing NELF were collected by centrifugation, resuspended in Lysis buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF, and 0.33 mg ml^{-1} benzamidine), flash-frozen in liquid nitrogen, and stored at -80°C until purification.

Human TFIIS was produced as a codon-optimized gBlock for *E. coli* expression (Integrated DNA Technologies). The gBlock was cloned into a modified pET28b vector bearing an N-terminal His6-MBP tag followed by a tobacco etch virus (TEV) protease cleavage site (1C vector, Addgene 29654). Two mutations were introduced by round-the-horn site-directed mutagenesis to prevent stimulation of RNA cleavage by Pol II (D282A/E283A). TFIIS was overexpressed in *E. coli* BL21 (DE3) RIL cells (Merck) grown in LB medium. Cells were grown at 37°C until reaching an optical density at 600 nm (OD_{600}) of around 0.6. The temperature was decreased to 18°C and protein expression was induced by adding 0.5 mM β -D-1-thiogalactopyranoside. Cells were grown for an additional 16 h at 18°C and were collected by centrifugation, resuspended in A800 (800 mM NaCl, 20 mM Tris-HCl pH 7.9, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF and 0.33 mg ml^{-1} benzamidine), flash-frozen in liquid nitrogen, and stored at -80°C . **Protein purification.** All protein purification steps were performed at 4°C unless otherwise stated. Pol II was isolated from *S. scrofa* thymus (obtained from E. Wolf, Ludwig Maximilian University of Munich) essentially as described⁵². A final size-exclusion step was performed using a Sephacryl S-300 16/60 column (GE Healthcare Life Sciences) equilibrated in 150 mM NaCl, 10 mM Na-HEPES pH 7.25, 10 μM ZnCl_2 , and 10 mM DTT. Peak fractions containing Pol II were concentrated in 100 kDa molecular weight cutoff (MWCO) Amicon Ultra Centrifugal Filters (Merck), aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C . The typical yield of Pol II from 1 kg of pig thymus was 5–7 mg.

DSIF was purified from 6–8 l of *E. coli*. Cell pellets were lysed by sonication, and cleared by centrifugation. The clarified lysate was filtered through a 0.8- μm syringe filter and applied to a 5-ml HisTrap HP column (GE Healthcare Life Sciences) equilibrated in Lysis 500 buffer. The column was washed with 10 column volumes of Lysis 500 buffer, followed by 3 column volumes of High Salt buffer (1,000 mM NaCl, 50 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 50 mM imidazole pH 8.0, and 1 mM DTT, 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF, and 0.33 mg ml^{-1} benzamidine). The column was washed with 3 column volumes of Lysis 500 buffer and the protein was eluted over a gradient with a buffer containing 500 mM NaCl, 50 mM Na-HEPES pH 7.4, 50 mM imidazole pH 8.0, 10% (v/v) glycerol, and 1 mM DTT, 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF, and 0.33 mg ml^{-1} benzamidine. Peak fractions containing DSIF were pooled, mixed with 3C protease and dialysed overnight in 7 kDa MWCO SnakeSkin dialysis tubing (Thermo Scientific) against Q buffer (300 mM NaCl, 50 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, and 1 mM DTT). The protein was applied to a tandem HisTrap (5 ml)/HiTrap Q (5 ml) column (GE Healthcare Life Sciences) equilibrated in Q buffer to remove the 3C protease, the His tag, uncleaved protein, and protein lacking the acidic N-terminal region of SPT5. The tandem column was then washed with 5 column volumes of Q buffer after which the HisTrap column was removed. The HiTrap Q column was developed over a gradient with High Salt buffer. Peak fractions were pooled and protein purity was assessed by SDS–PAGE and Coomassie staining. Pure DSIF was concentrated with 50 kDa MWCO Amicon Ultra Centrifugal Filters (Merck) and applied to a HiLoad S200 16/600 pg column equilibrated in 500 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, and 1 mM DTT. Protein purity was assessed by SDS–PAGE and Coomassie staining. Pure fractions with full-length SPT5 were concentrated with 50 kDa MWCO Amicon Ultra Centrifugal Filters (Merck).

Protein concentration was determined by measuring absorption at 280 nm and using the predicted extinction coefficient for the complex. Protein was aliquoted, flash-frozen, and stored at -80°C . NELF was purified as previously described²¹.

TFIIS was purified from 6 l of *E. coli* BL21 (DE3) RIL cells. Cell pellets were lysed by sonication and cleared by centrifugation. Clarified lysates were applied to 5 ml HisTrap columns equilibrated in A800. The column was washed with A800 and A400 (400 mM NaCl, 20 mM Tris-HCl pH 7.9, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 $\mu\text{g ml}^{-1}$ leupeptin, 1.37 $\mu\text{g ml}^{-1}$ pepstatin A, 0.17 mg ml^{-1} PMSF, and 0.33 mg ml^{-1} benzamidine) until no additional absorbance was detected at 280 nm. The protein was eluted from the nickel column by a gradient over 6 columns with buffer B400 (A400 with 500 mM imidazole pH 8.0). Peak fractions were assessed by SDS–PAGE followed by Coomassie staining. Fractions corresponding to TFIIS were pooled and mixed with TEV protease, and dialysed overnight against buffer A400 in SnakeSkin dialysis tubing (7 kDa MWCO). The protein was removed from dialysis and applied to a 5 ml HisTrap column equilibrated in A400 to remove TEV protease, uncleaved protein, and the His6-MBP tag. The flow through was collected, concentrated in a 10-kDa MWCO Amicon Ultra Centrifugal Filter (Merck), and applied to a HiLoad S75 16/1600 pg column equilibrated in 400 mM NaCl, 20 mM Tris-HCl pH 7.9, 10% (v/v) glycerol, and 1 mM DTT. Protein purity was assessed by SDS–PAGE followed by Coomassie staining. Peak fractions were concentrated in 10 kDa MWCO Amicon Ultra Centrifugal Filter (Merck). The protein was stored in buffer containing 400 mM NaCl, 20 mM Tris-HCl pH 7.9, 30% (v/v) glycerol, and 1 mM DTT. The protein was aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C .

RNA extension assays. All oligos were purchased from Integrated DNA Technologies, resuspended in water (100 μM), flash-frozen in liquid nitrogen, and stored at -80°C . Transcription assays were performed with perfectly complementary scaffolds. A previously described pausing sequence²⁴, the ‘pause assay scaffold’ with the following sequence was used for experiments in Figs. 1 and 6: template DNA 5′-Biotin-TTT TTC CAC TGG AAG ATC TGA ATT TAC GGG CGC AAC TAT GCC GGA CGT ACT GAC C-3′, non-template DNA 5′-GGT CAG TAC GTC CGG CAT AGT TGC GCC CGT AAA TTC AGA TCT TCC AGT GG-3′, RNA 5′-6-FAM-UUU UUU GGC AUA GUU-3′. The scaffold contains 13 nucleotides of upstream DNA, 28 nucleotides of downstream DNA, a 9-base-pair DNA–RNA hybrid, and 6 nucleotides of exiting RNA bearing a 5′-6 FAM label (Extended Data Fig. 1). RNA and template DNA were mixed in equimolar ratios and were annealed by incubating the nucleic acids at 95°C for 5 min and then decreasing the temperature by $1^{\circ}\text{C min}^{-1}$ steps to a final temperature of 30°C in a thermocycler in a buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl_2 , and 10% (v/v) glycerol. All concentrations refer to the final concentrations used in the assay. *S. scrofa* Pol II (75 nM) and the RNA–DNA template hybrid (50 nM) were incubated for 10 min at 30°C , shaking at 300 rpm. The non-template DNA (50 nM) was added and the reactions were incubated for another 10 min. The reactions were then diluted to achieve final assay conditions of 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl_2 , 4% (v/v) glycerol, and 1 mM DTT and were again incubated for 10 min. Factors were diluted in protein dilution buffer (150 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol and 1 mM DTT) and added to Pol II ECs at a concentration of 237 nM. Transcription reactions were initiated by adding GTP and CTP (10 μM) to permit elongation to position +7. Reactions (10 μl) were quenched after 0–10 min in 10 μl 2x Stop buffer (6.4 M urea, 50 mM EDTA pH 8.0, 1x TBE buffer). Samples were treated with 4 μg of proteinase K for 30 min at 30°C (New England Biolabs) and were separated by denaturing gel electrophoresis (8 μl of sample applied to an 8 M urea, 1x TBE, 20% Bis-Tris acrylamide 19:1 gel run in 0.5x TBE buffer at 300V for 90 min). Products were visualized using the 6-FAM label and a Typhoon 9500 FLA Imager (GE Healthcare Life Sciences).

RNA extension experiments performed on the HIV-1 transcription scaffold (Extended Data Fig. 2a) were essentially performed as above with minor modifications. ECs were assembled on nucleic acid scaffolds bearing the following sequences: template DNA 5′-Biotin-TTT TCG GGC ACA CAC TAC GTC GAC GCA AGC TTT ATT GAG GCT TAA GCA GTG GGT TCC CTA GTT AAA GGT ACT AGT GTA C-3′, non-template DNA 5′-GTA CAC TAG TAC CTT TAA CTA GGG AAC CCA CTG CTT AAG CCT CAA TAA AGC TTG CGT CGA CGT AGT GTG TGC CCG-3′, RNA 5′-6-FAM-ACC AGA UCU GAG CCU GGG AGC UCU CUG GCU AAC UAG GG-3′. The scaffold contains 15 base pairs of upstream DNA, 51 base pairs of downstream DNA, a 9-base-pair RNA–DNA hybrid, and 29 bases of exiting RNA including the TAR element. DSIF and NELF variants were added at a final concentration of 300 nM. NTPs (ATP, UTP, CTP and GTP) were added at a final concentration of 0.5 mM. RNA products were separated by denaturing gel electrophoresis (8 μl of sample applied to an 8 M urea, 1x TBE, 15% Bis-Tris acrylamide 19:1 gel run in 0.5x TBE buffer at 300 V for 60 min).

Gel images were quantified using ImageJ version 1.48v⁵³. The integrated density of the elongated product was measured using a box size of 0.35×0.15 cm. All integrated density values were normalized by subtracting the background integrated density from each elongated product. Graphs were prepared in GraphPad Prism version 6. Each bar or point represents the mean intensity from three individual replicates. Error bars reflect the standard deviation between the replicates. Source data for all gel quantification can be found in Supplementary Table 6.

We observe extension from a fraction of the input RNA molecules. We attribute this to inefficient EC assembly on the perfectly complementary scaffolds. It was previously shown that only 10–50% of yeast Pol II molecules successfully assemble on perfectly complementary scaffolds^{54–56} owing to non-template DNA displacement of the RNA primer⁵⁶. Others have resolved the problem of displaced RNA primer by incorporating radioactive NTPs or by immobilizing non-template DNA containing complexes on beads. We chose to perform RNA extension experiments in bulk with a fluorescently labelled RNA to maintain consistent Pol II concentrations across experiments and reproducibility in time-course experiments.

Sample preparation for cryo-EM. ECs were formed on a bubble scaffold with the following nucleic acid sequence: template DNA 5'-GGC AAG CTT TAT TGA GGC TTA AGC AGT GGG TTC CAG GTA CTA GTG TAC-3', non-template DNA 5'-GTA CAC TAG TAC CTA CTC GAG TGA GCT TAA GCC TCA ATA AAG CTT GCC-3', and RNA 5'-6-FAM-ACC AGA UCU GAG CCU GGG AGC UCU CUG GCU AAC UAG GGA ACC CAC U-3'. The scaffold contains 10 bp RNA–DNA hybrid, 36 nucleotides of exiting RNA, 10-nucleotide bubble, 14 nucleotides upstream DNA and 24 nucleotides of downstream DNA. Pol II ECs were formed as described for the transcription assays (112 pmol final Pol II, 168 pmol RNA–DNA template, 300 pmol non-template DNA). DSIF and NELF were added in a fourfold molar excess relative to Pol II in a final buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl₂, 1 mM DTT, and 4% (v/v) glycerol. The sample was incubated for 30 min at 30 °C and applied to a Superose 6 increase 3.2/300 column equilibrated in complex buffer at 4 °C (100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 3 mM MgCl₂, and 1 mM DTT). Peak fractions were analysed by SDS–PAGE followed by Coomassie staining.

The peak fraction corresponding to the complex was crosslinked with 0.1% (v/v) glutaraldehyde for 10 min on ice and quenched with 8 mM aspartate and 2 mM lysine. The crosslinked sample was dialysed against a buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 20 mM Tris–HCl pH 7.5, 1 mM DTT, and 3 mM MgCl₂, in 20 kDa MWCO Slide-A-Lyzer MINI Dialysis Units for 6 h at 4 °C. Dialysed sample at a final concentration of 170–200 nM was applied to R2/2 gold grids (Quantifoil). The grids were glow-discharged for 45 s before applying 2 μ l of sample to each side of the grid (4 μ l total). After incubation for 10 s and blotting for 8.5 s, the grid was vitrified by plunging it into liquid ethane with a Vitrobot Mark IV (FEI Company) operated at 4 °C and 100% humidity.

Cryo-EM data collection and data processing. Cryo-EM data of the PEC were collected on a FEI Titan Krios II transmission electron microscope operated at 300 keV. A K2 summit direct detector (Gatan) with a GIF quantum energy filter (Gatan) was operated with a slit width of 20 eV. Automated data acquisition was performed with FEI EPU software at a nominal magnification of 165,000 \times , corresponding to a pixel size of 0.81 Å per pixel. Image stacks of 36 frames were collected over 9 s in counting mode. The dose rate was 5.1 e[−] per Å² per s for a total dose of 45.9 e[−] per Å². A total of 11,740 image stacks were collected.

Frames were stacked, contrast-transfer-function corrected, and dose-weighted using Warp⁵⁷. The data were binned to a pixel size of 1.2277 Å per pixel. Image processing was performed with RELION 2.1^{58,59}. Particles were auto-picked using 20-Å low-pass filtered projections of an initial reconstruction yielding 2,347,915 particle images. Particles were extracted using a box size of 256² pixels, and normalized. The dataset was segmented in three batches. Each batch was subsequently screened using iterative rounds of reference-free 2D classification. A cryo-EM reconstruction of the EC–DSIF complex (EMDB: 3819)¹⁹ was low-pass-filtered to 50 Å, used for 3D refinement and hierarchical 3D-classification with image alignment. We obtained an EC–DSIF bound class which contained 479,365 particles and resulted in a reconstruction at 2.9 Å resolution indicating the high quality of the raw data (Extended Data Figs. 3, 4). The best-resolved NELF-bound classes from each batch were selected and combined resulting in 162,269 particles. The combined particles were subjected to 3D refinement using a 30 Å low-pass-filtered map from a previous 3D-refinement resulting in a reconstruction with a resolution of 3.2 Å (map 1). Some domains were not well resolved in the reconstruction, so 3D classifications without image alignment were performed around the regions of interest by applying soft masks. Masks were generated in Chimera⁶⁰ and RELION 2.1 around the NELF-A–NELF-C dimer (map 2, 3.4 Å), NELF-B (map 3, 3.4 Å), upstream DNA, NGN and KOW1 (map 4, 3.9 Å) subunits/domains, and the stalk and KOW2–3 (map 5, 3.3 Å). Particles containing the desired densities were subjected to global 3D refinement. To further improve densities, a masked refinement was used for NELF-B. The masked refinement was performed

by using final alignments from previous global refinements with local searches and by applying soft masks to the region of interest. The masked refinement performed on NELF-ABC resulted in a resolution of 3.9 Å with an applied *B*-factor of -160.984 Å². Post-processing of refined models was performed using automatic *B*-factor determination in RELION and reported resolutions are based on the gold-standard Fourier shell correlation 0.143 criterion⁶¹ (applied *B*-factors (Å²): map 1: -65 , map 2: -65 , map 3: -103 , map 4: -71 , map 5: -68). Local resolution estimates were determined using the built-in local resolution estimation tool of RELION using the previously estimated *B*-factors⁶².

Model building. The structure of the PEC was built by first placing the structure of RNA polymerase II from the activated elongation complex⁵⁰ (EC*) manually into the density. Adjustments were made to the protein sequence, DNA sequence, and positioning of the upstream DNA in Coot⁶³. The human RPB4-7 crystal structure (PDB ID: 2C35)⁶⁴ was placed into map 5 using Chimera.

Human DSIF from a previously solved cryo-EM structure (PDB ID: 5OIK)¹⁹ was subdivided into five regions for modelling, corresponding to the SPT5 NGN and SPT4, KOW1, KOW2–KOW3, KOWx–KOW4 and KOW5. KOW5 was placed into the globally refined map 1. SPT4, NGN domain and KOW1 domain were placed in map 4 by rigid-body fitting in Chimera. The KOW2–KOW3 and KOWx–KOW4 were placed in map 5 by rigid-body fitting in Chimera.

Densities corresponding to each NELF subunit were observed. To generate a model for the NELF complex, the known crystal structure of the NELF-A–NELF-C dimer (PDB ID: 5L3X)²¹ was flexibly fit into the globally refined density of map 1 using VMD and MDFF⁶⁵. The model was manually adjusted in Coot. The N-terminal part of NELF-C (51–185) was built de novo in Coot in the globally refined map 1 using secondary structure prediction and the well-resolved helical densities that allowed placement of helices into the density.

A model for NELF-B was generated using de novo and homology modelling. Secondary structure predictions from SABLE⁶⁶ and PSIPRED⁶⁷ were used to assist de novo modelling. Clear helical densities are observed for all helices predicted by SABLE and PSIPRED. Alpha helices were generated using Coot and manually fitted into the density. Linkers between the helices were modelled where clear density was visible. A Robetta model of NELF-B (residues 438–548) was fit into map 3⁶⁸. Crosslinking restraints and densities from bulky residues such as Arg and Tyr were used as additional sequence markers for NELF-B and the N terminus of NELF-C. The staircase domain (NELF-B 153–365) shares modest structural homology with yeast exportin-1 alpha (r.m.s.d. 13.95 Å, PDB: 4HAX, Chain C, residues 388–638)^{69,70}. NELF-B 408–548 is composed of 4 HEAT repeats^{69,71}.

NELF-E is the least well resolved subunit in our structure. We observe an additional helix immediately adjacent to the HEAT domain of NELF-B that cannot be assigned to NELF-B. Crosslinking data and secondary structure predictions assigned this helix to NELF-E residues 10–34. This assignment is consistent with biochemical experiments⁸.

The model was manually adjusted in Coot⁶³ and refined with phenix.real_space_refine against a sharpened version of Map 3. The final model has 94.10% of residues in most-favoured regions of the Ramachandran plot according to Molprobity⁷². The structure has a Molprobity score of 1.79. Figures were generated in PyMOL (Schrödinger LLC, version 1.8.6.0) and UCSF Chimera (version 1.10.2).

Analytical gel filtration. Pol II ECs were formed on the same scaffold as was used for cryo-EM (25 pmol final Pol II, 50 pmol RNA–DNA template, 100 pmol NT DNA). DSIF and NELF were added in a 3-molar excess relative to Pol II, whereas TFIIIS was added in a 11-molar excess in a final buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl₂, 1 mM DTT, and 4% (v/v) glycerol. Reactions were incubated for 30 min at 30 °C. Samples were applied to a Superose 6 increase 3.2/300 column equilibrated in complex buffer (100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 3 mM MgCl₂, and 1 mM DTT). Peak fractions were analysed by SDS–PAGE followed by Coomassie staining.

Sample preparation for crosslinking mass spectrometry. Samples for crosslinking mass spectrometry were performed essentially in the same way as those for cryo-EM. A nucleic acid scaffold that differs by two nucleotide bases from the scaffold used for cryo-EM was used (template DNA 5'-GGC AAG CTT TAT TGA GGC TTA AGC AGT GGG TTC AAG GTA CTA GTG TAC-3', non-template DNA 5'-GTA CAC TAG TAC CTA CTC GAG TGA CCT TAA GCC TCA ATA AAG CTT GCC-3', RNA sequence is identical). Fractions containing the PEC were pooled and mixed with 2 mM of bis(sulfosuccinimidyl)suberate (BS3) dissolved in complex buffer (No Weigh Format, ThermoFisher Scientific). The protein was incubated for 30 min at 30 °C. The crosslinking reaction was quenched by adding 100 mM Tris–HCl pH 7.5 and 20 mM ammonium bicarbonate (final concentrations). The quenching reaction was incubated for 15 min at 30 °C. The protein was precipitated with 300 mM sodium acetate pH 5.2 and 4 volumes of acetone and incubated overnight at -20 °C, pelleted by centrifugation, briefly dried, and resuspended in 4 M urea and 50 mM ammonium bicarbonate.

Crosslinking mass spectrometry. Crosslinked proteins were reduced with 10 mM DTT for one hour at room temperature. Alkylation was performed by adding

iodoacetamide to a final concentration of 40 mM, incubated for 30 min in the dark at room temperature. After dilution to 1 M urea with 50 mM ammonium bicarbonate (pH 8.0), the crosslinked protein complex was digested with trypsin in a 1:50 enzyme-to-protein ratio at 37 °C overnight. Peptides were acidified with trifluoroacetic acid (TFA) to a final concentration of 0.5% (v/v), desalted on MicroSpin columns (Harvard Apparatus) following the manufacturer's instructions and vacuum-dried. Dried peptides were dissolved in 50 µl 30% acetonitrile/0.1% TFA and peptide size-exclusion (pSEC, Superdex Peptide 3.2/300 column on an ÄKTAmicro system, GE Healthcare) was performed to enrich for crosslinked peptides at a flow rate of 50 µl min⁻¹. Fractions of 50 µl were collected. Fractions containing the crosslinked peptides (1–1.7 ml) were vacuum-dried and dissolved in 2% acetonitrile/0.05% TFA (v/v) for analysis by liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS).

Crosslinked peptides derived from pSEC were analysed as technical duplicates on an Orbitrap Fusion and Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Scientific), respectively, coupled to a Dionex UltiMate 3000 UHPLC system (Thermo Scientific) equipped with an in-house-packed C₁₈ column (ReproSil-Pur 120 C18-AQ, 1.9 µm pore size, 75 µm inner diameter, 30 cm length, Dr. Maisch GmbH). Samples were separated applying the following 58-min gradient: mobile phase A consisted of 0.1% formic acid (v/v), mobile phase B consisted of 80% acetonitrile/0.08% formic acid (v/v). The gradient started at 5% B, increasing to 8% B on Fusion and 15% on Fusion Lumos, respectively, within 3 min, followed by 8–42% B and 15–46% B within 43 min accordingly, then keeping B constant at 90% for 6 min. After each gradient the column was again equilibrated to 5% B for 6 min. The flow rate was set to 300 nl min⁻¹. MS1 spectra were acquired with a resolution of 120,000 in the Orbitrap covering a mass range of 380–1,580 m/z. The injection time was set to 60 ms and the automatic gain control target to 5×10^5 . Dynamic exclusion covered 10 s. Only precursors with a charge state of 3–8 were included. MS2 spectra were recorded with a resolution of 30,000 in the Orbitrap, injection time was set to 128 ms, automatic gain control target to 5×10^4 and the isolation window to 1.6 m/z. Fragmentation was enforced by higher-energy collisional dissociation at 30%.

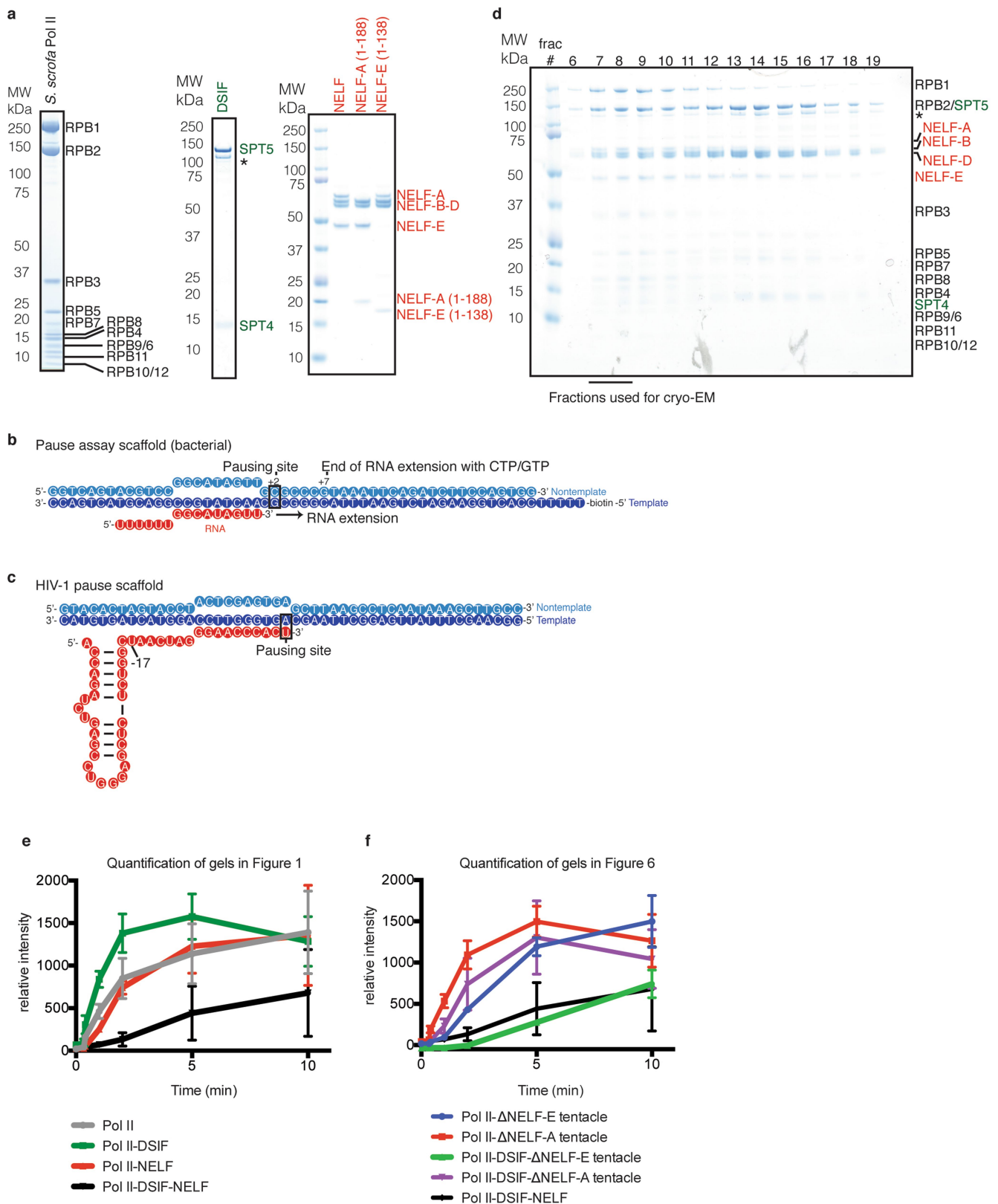
Raw files were converted to mgf format using ProteomeDiscoverer 1.4 (Thermo Scientific, signal-to-noise ratio 1.5, 1,000–10,000 Da precursor mass). For identification of crosslinked peptides, files were analysed by pLink (v. 1.23, pFind group⁷³ using BS3 as crosslinker and trypsin as digestion enzyme with maximal two missed cleavage sites. Carbamidomethylation of cysteines was set as a fixed modification, oxidation of methionines as a variable modification. Searches were conducted in combinatorial mode with a precursor mass tolerance of 5 Da and a fragment ion mass tolerance of 20 p.p.m. The used database contained all proteins within the complex. The false discovery rate was set to 0.01. Results were filtered by applying a precursor mass accuracy of ± 10 p.p.m. Spectra of both technical duplicates were combined. Crosslinking figures were made with XiNet⁷⁴ and the Xlink Analyzer plugin in Chimera^{60,75}. Distances between structured regions were calculated with Xlink Analyzer with a cutoff score of 5.

A total of 874 unique crosslinks were obtained of which 354 could be mapped onto our structure. 261 are located within the 30 Å distance permitted by BS3 whereas the remaining 93 crosslinks primarily lie in flexible regions of NELF. The NELF crosslinks are highly similar to those obtained with the isolated NELF complex²¹.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The electron density reconstructions and final PEC model have been deposited in the Electron Microscopy Data Bank (EMDB) under accession codes EMD-0038, EMD-0039, EMD-0040, EMD-0041 and EMD-0042, and with the Protein Data Bank (PDB) under accession code 6GML. Source data for Figs. 1a, b and 6c, Extended Data Figs. 1a, d–f, 2b–e can be found in Supplementary Fig. 1 and Supplementary Table 6.

51. Gradiš, S. D. et al. MacroBac: new technologies for robust and efficient large-scale production of recombinant multiprotein complexes. *Methods Enzymol.* **592**, 1–26 (2017).
52. Hu, X. et al. A mediator-responsive form of metazoan RNA polymerase II. *Proc. Natl Acad. Sci. USA* **103**, 9506–9511 (2006).
53. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
54. Xu, J. et al. Structural basis for the initiation of eukaryotic transcription-coupled DNA repair. *Nature* **551**, 653–657 (2017).
55. Kireeva, M. L., Komissarova, N., Waugh, D. S. & Kashlev, M. The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J. Biol. Chem.* **275**, 6530–6536 (2000).
56. Komissarova, N., Kireeva, M. L., Becker, J., Sidorenkov, I. & Kashlev, M. Engineering of elongation complexes of bacterial and yeast RNA polymerases. *Methods Enzymol.* **371**, 233–251 (2003).
57. Tegunov, D. & Cramer, P. Real-time cryo-EM data pre-processing with Warp. Preprint at <https://www.biorxiv.org/content/early/2018/06/14/338558> (2018).
58. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
59. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
60. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
61. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
62. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D* **73**, 496–502 (2017).
63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
64. Meka, H., Werner, F., Cordell, S. C., Onesti, S. & Brick, P. Crystal structure and RNA binding of the Rpb4/Rpb7 subunits of human RNA polymerase II. *Nucleic Acids Res.* **33**, 6435–6444 (2005).
65. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
66. Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59**, 467–475 (2005).
67. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
68. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526–W531 (2004).
69. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 (2010).
70. Sun, Q. et al. Nuclear export inhibition through covalent conjugation and hydrolysis of Leptomycin B by CRM1. *Proc. Natl Acad. Sci. USA* **110**, 1303–1308 (2013).
71. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18 (2001).
72. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
73. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
74. Combe, C. W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol. Cell. Proteomics* **14**, 1137–1147 (2015).
75. Kosinski, J. et al. Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J. Struct. Biol.* **189**, 177–183 (2015).
76. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
77. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

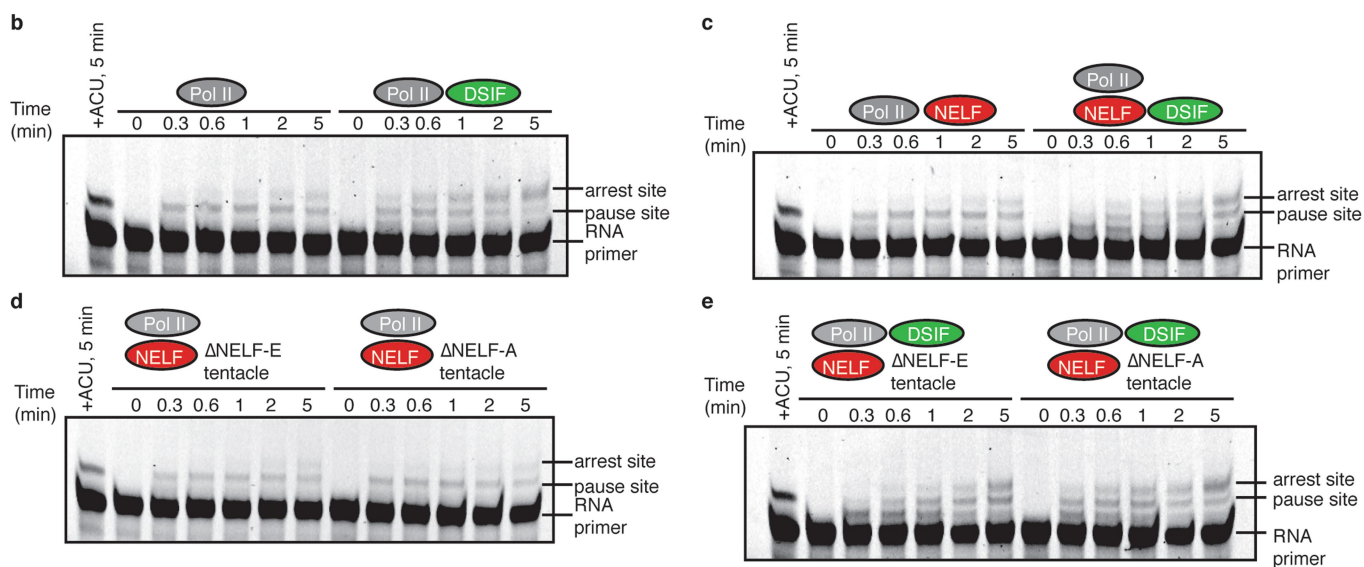
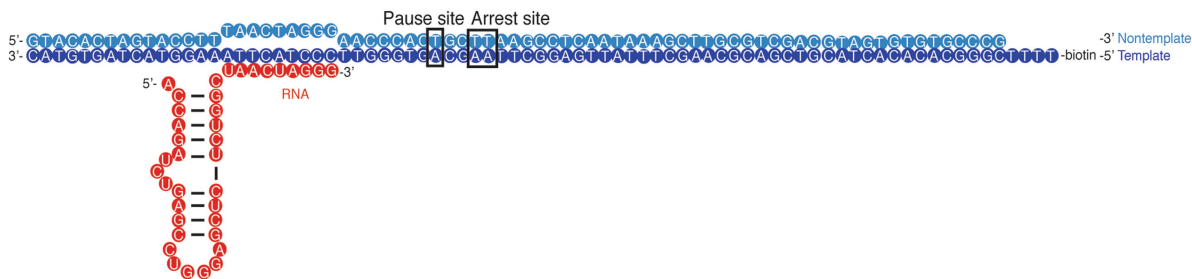


Extended Data Fig. 1 | Protein preparation and nucleic acid scaffold design. **a**, Quality of purified proteins used in this study. Purified proteins (0.9 μ g) were run on 4–12% SDS–PAGE and stained with Coomassie blue. An asterisk demarcates SPT5 lacking an N-terminal region. **b**, Nucleic acid scaffold used for RNA extension assays, the ‘pause assay scaffold’. Template DNA is coloured in dark blue, non-template DNA is in light blue, and RNA is in red. **c**, Nucleic acid scaffold used for binding experiments and for cryo-EM analysis, the ‘HIV-1 pause scaffold’. Colours are the same as in **b**. **d**, SDS–PAGE analysis of fractions obtained from size-exclusion

chromatography. The fractions used for cryo-EM analysis are marked. **e**, Quantification of the RNA extension assays shown in Fig. 1. The amount of elongated product was measured for each time point. Points are the mean of three independent experiments and error bars represent the standard deviation between experiments. **f**, Quantification of the RNA extension assays shown in Fig. 6. The amount of elongated product was measured for each time point. Points are the mean of three independent experiments and error bars represent the standard deviation between experiments.

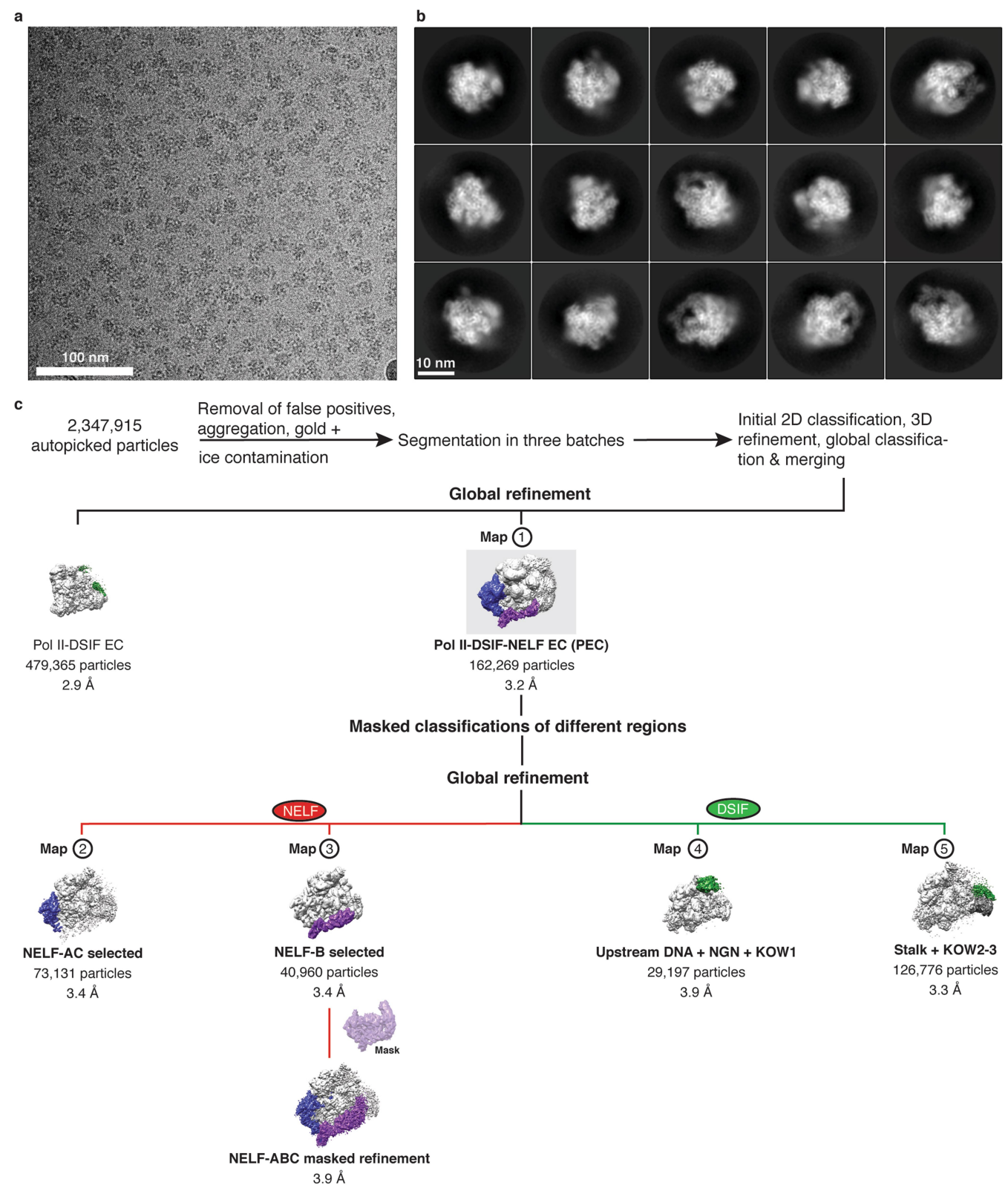
a

HIV-1 transcription scaffold



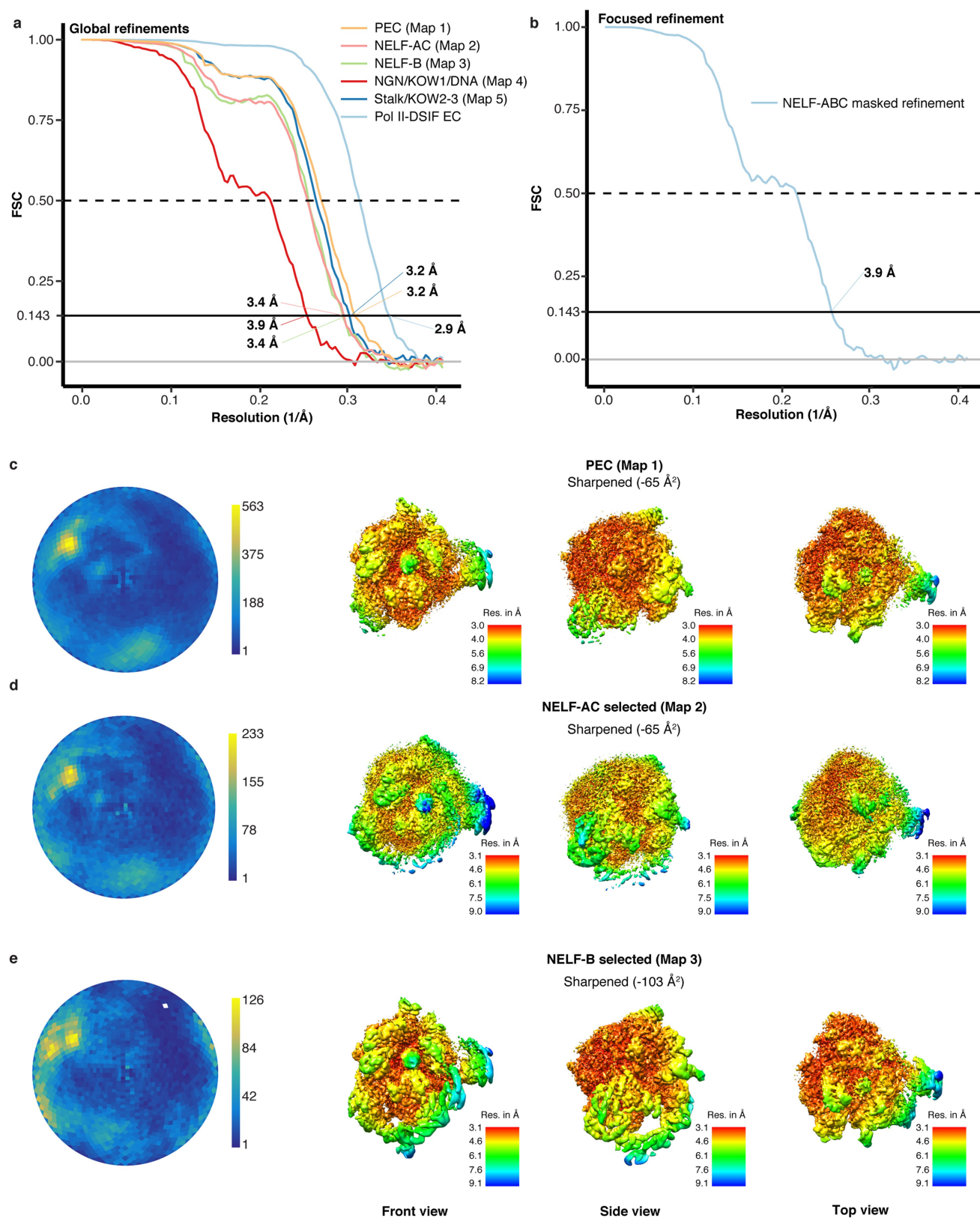
Extended Data Fig. 2 | RNA extension assays on HIV-1 nucleic acid scaffold. **a**, HIV-1 nucleic acid scaffold used for RNA extension assays. The sequence is slightly altered from that used for cryo-EM to allow extension for eight bases before pausing. Known pause and arrest sites are marked on the sequence. **b–e**, Pol II ECs (75 nM) were reconstituted on the HIV-1 transcription scaffold (50 nM). A single reaction was incubated with ATP, CTP and UTP (0.5 mM) for 5 min to indicate the pause site

(far right lane). Buffer (**b**), DSIF (**b**), NELF (**c**), DSIF and NELF (**c**), NELF tentacle mutants (**d**), or DSIF and NELF tentacle mutants (**e**) (300 nM) were incubated with the Pol II EC. NTPs were added (0.5 mM) and aliquots were taken at specific time points. Only a fraction of the starting RNA is successfully elongated owing to incomplete EC formation (see Methods for more information).



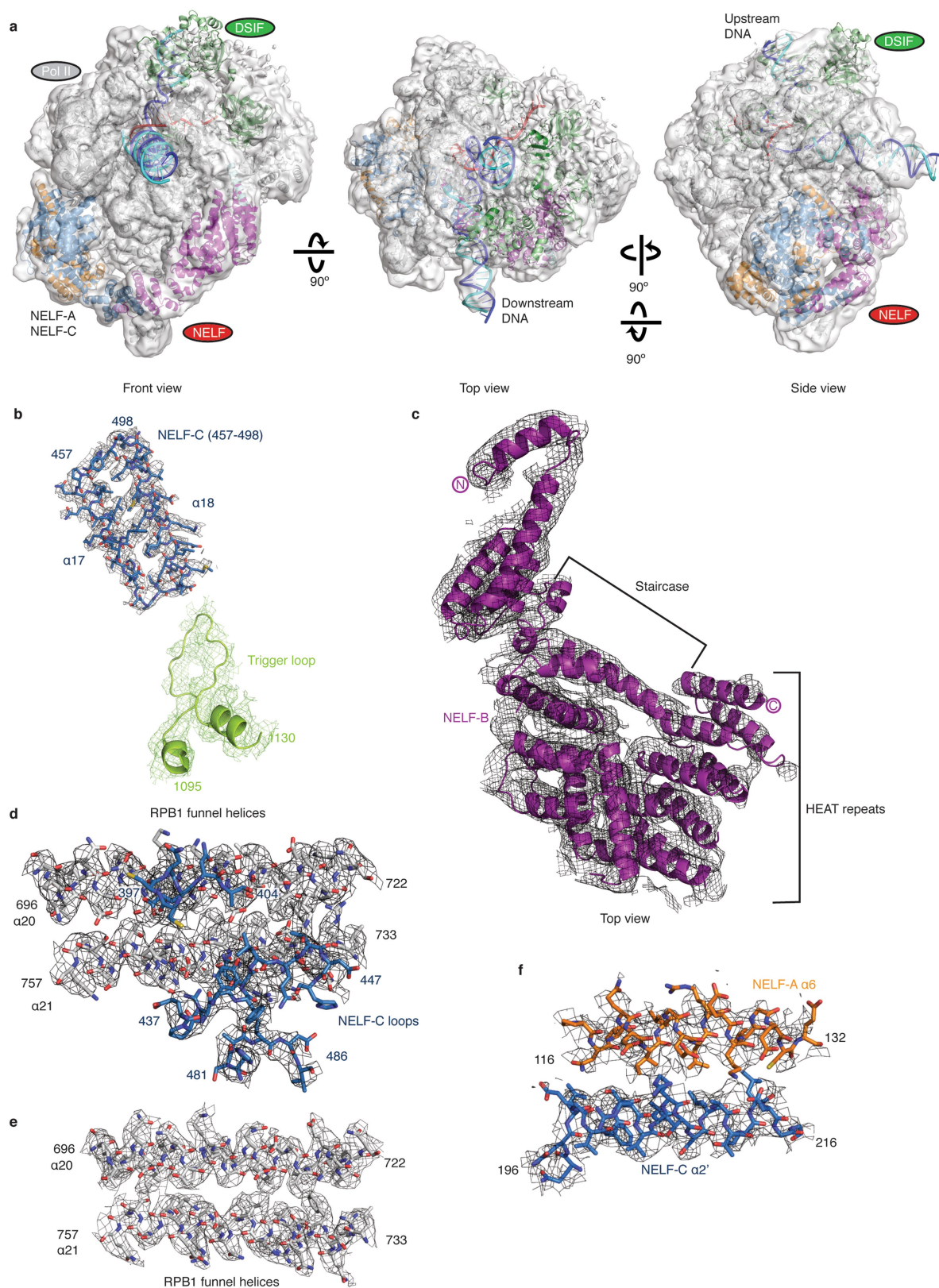
Extended Data Fig. 3 | Cryo-EM data collection and processing.
a, Representative micrograph of data collection for the PEC, shown at a defocus of 2.5 μm . The micrograph is representative of 11,740

micrographs. **b**, Representative 2D classes of PEC particles.
c, Classification tree for data processing. The numbers used to identify each map are shown above the corresponding map.



Extended Data Fig. 4 | Quality of cryo-EM data. **a, b**, Estimation of average resolution, showing global (**a**) and focused (**b**) refinement. The lines indicate the FSC between the half maps of the reconstruction. FSC curves are shown for each map. **c–e**, Angular distribution of particles from overall refinements and local resolution of selected refinements for the PEC (map 1) (**c**), NELF-A–NELF-C selected (map 2) (**d**) and

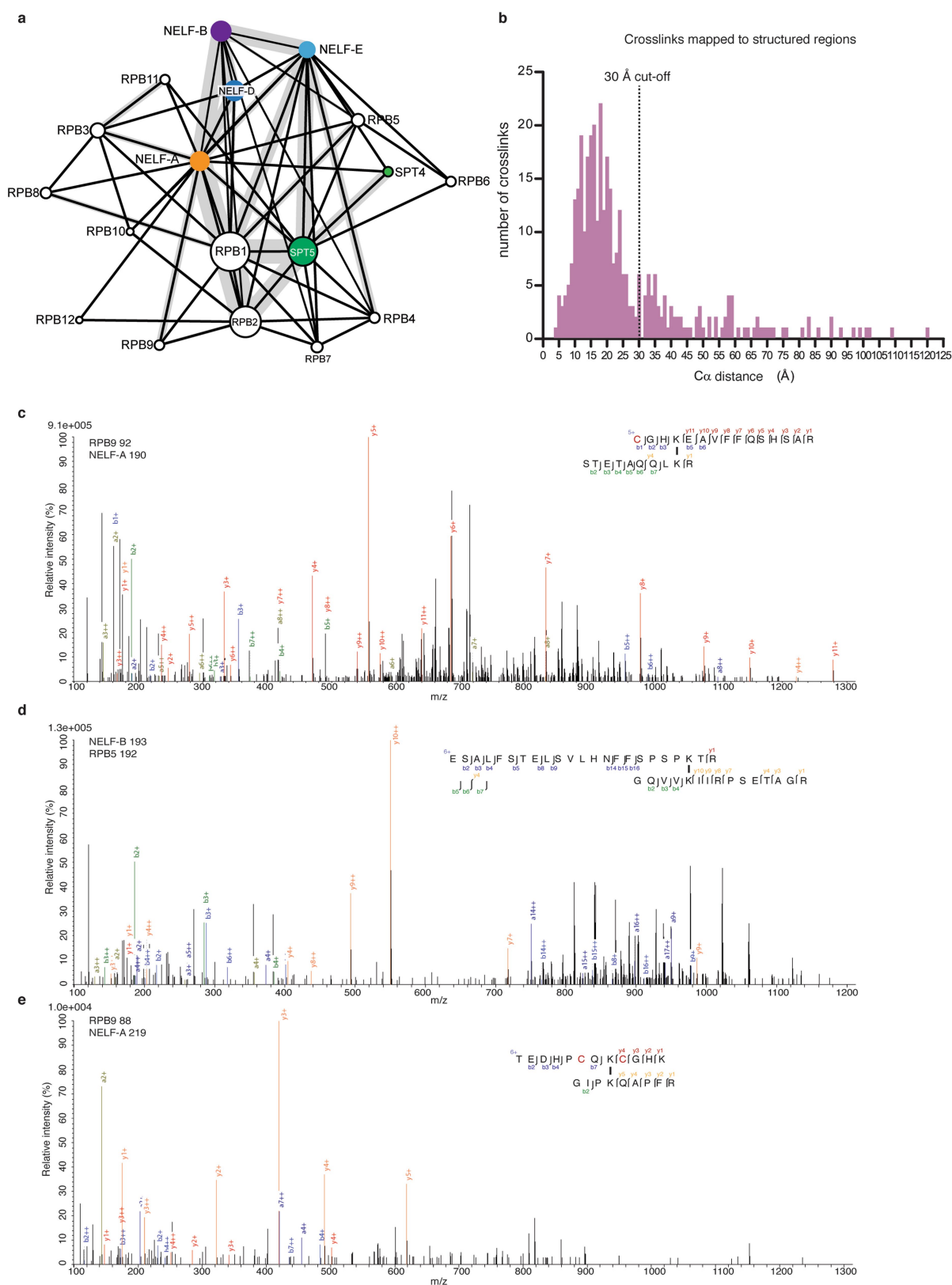
NELF-B selected (map 3) (**e**). Shading from blue to yellow indicates the number of particles at a given orientation. Reconstructions coloured by local resolution. Shading from red to blue indicates the local resolution according to the accompanying colour gradient. Absolute values are indicated. B-factors were used as indicated.



Extended Data Fig. 5 | Fit of PEC structure in representative densities.

a, PEC structure fit in electron density contoured to 6 Å from map 3. Front, top, and side views are shown. **b–f**, Electron density for various elements of the PEC structure shown as meshes. **b**, A loop connecting

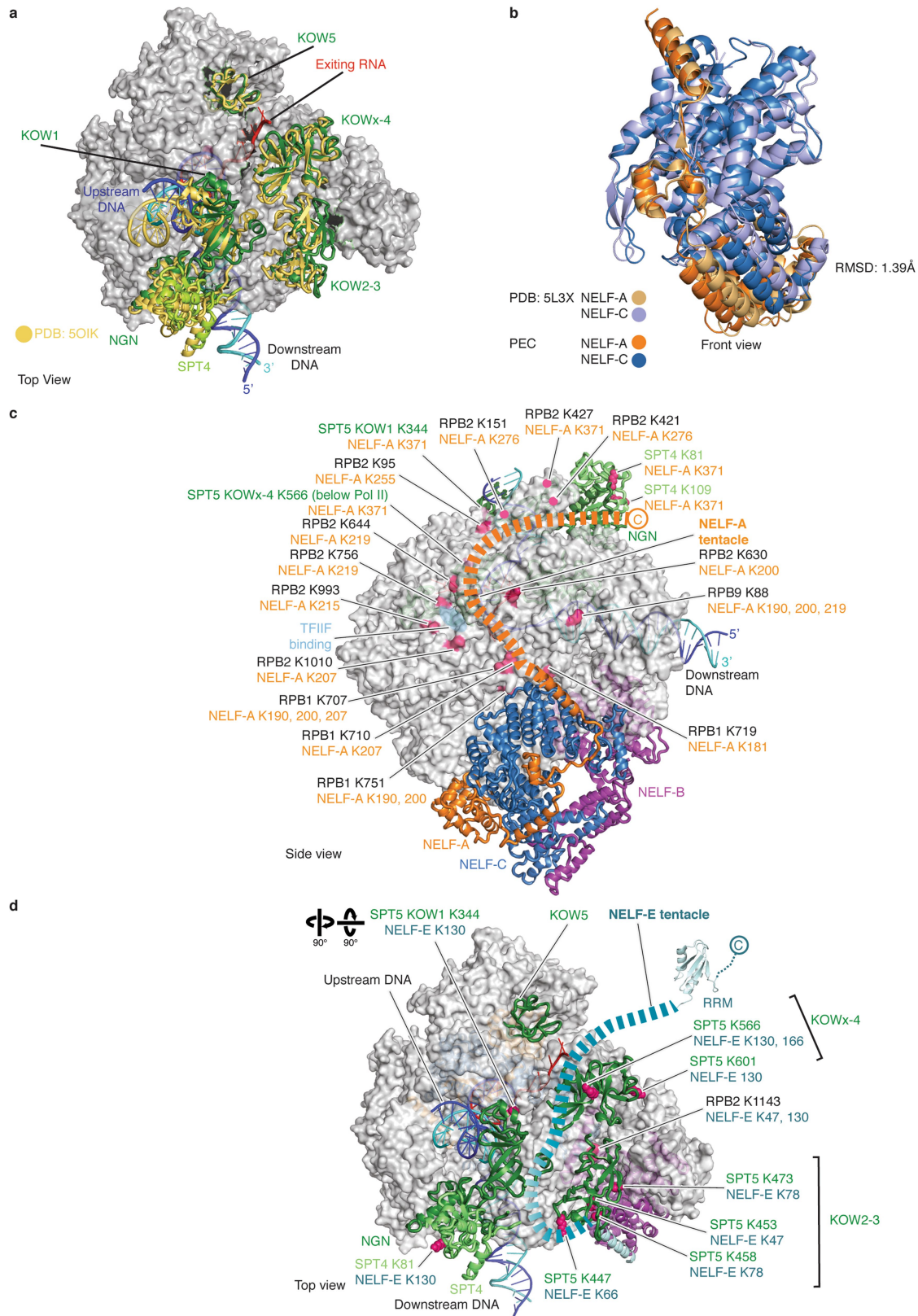
NELF-C helices 17 and 18 (map 3, grey mesh) contacts the trigger loop (map 2, lime-green mesh). **c**, NELF-B (map 3). **d**, NELF-C contacts the RPB1 funnel helices ($\alpha 20$, $\alpha 21$). **e**, Funnel helices ($\alpha 20$, $\alpha 21$). **f**, The NELF-A–NELF-C interaction (A- $\alpha 6$, C- $\alpha 2'$).



Extended Data Fig. 6 | Crosslinking mass spectrometry analysis.

a, Overview of PEC crosslinks obtained with BS3. Subunits coloured as in Fig. 1. The thickness of the grey line connecting subunits signifies the number of crosslinks obtained between subunits. **b**, Histogram of unique crosslinks that were mapped onto our structure. Distances are measured between $C\alpha$ pairs using Xlink analyser⁷⁵ for crosslinks with a score greater than 5. The number of unique crosslinks detected at each distance is indicated. A dotted black line marks the 30 Å distance cut-off for BS3.

c–e, Representative spectra from crosslinking mass spectrometry experiments. Blue, red and dark blue correspond to b-, y-, and a-ions of peptide A, respectively. Green, orange and dark green correspond to b-, y-, and a-ions of peptide B. Black bars drawn between lysines indicate crosslinking sites. Red highlighted 'C' represents carbamido-methylated cysteine residues. Relative intensity of m/z is plotted. Spectra are representative of one biological and two technical replicates.



Extended Data Fig. 7 | Comparison of previous structures to the PEC. **a**, The PEC and Pol II-DSIF EC structures were aligned by their Pol II cores. Slight differences are observed in DSIF bound to the PEC (green) in comparison to the Pol II-DSIF EC¹⁹ (yellow). **b**, The previously solved NELF-A-NELF-C dimerization crystal structure²¹ (PDB ID: 5L3X) and the NELF-A-NELF-C dimerization domain from the PEC cryo-EM

structure were aligned on the NELF-C subunit. The NELF-A-NELF-C dimer widens when bound to Pol II (r.m.s.d. 1.39 Å). **c**, NELF-A tentacle crosslinks mapped onto the PEC. NELF-A and corresponding Pol II or DSIF residues are indicated. Related to Fig. 6. **d**, NELF-E tentacle crosslinks mapped onto the PEC. NELF-E and corresponding Pol II or DSIF residues are indicated. Related to Fig. 6.

a

Conservation of RPB1 shelf module residues that interact with NELF-C

Conservation of the NELF core residues that interact with RNA Pol II

		708	712			743	747																																																			
Organisms encoding NELF	<i>H. sapiens</i>	700	QD	I	Q	N	T	I	K	K	A	K	A	K	D	V	I	E	V	I	E	K	A	H	N	N	E	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	R	I	L	N	D	A	R	D	K	T	G	S	754
	<i>M. musculus</i>	700	QD	I	Q	N	T	I	K	K	A	K	A	K	D	V	I	E	V	I	E	K	A	H	N	N	E	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	R	I	L	N	D	A	R	D	K	T	G	S	754
	<i>X. laevis</i>	698	QD	I	Q	N	T	I	K	K	A	K	A	K	D	V	I	E	V	I	E	K	A	H	N	N	E	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	R	I	L	N	D	A	R	D	K	T	G	S	752
	<i>D. rerio</i>	666	QD	I	Q	N	T	I	K	K	A	K	A	K	D	V	I	E	V	I	E	K	A	H	N	N	E	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	R	I	L	N	D	A	R	D	K	T	G	S	726
	<i>D. melanogaster</i>	692	N	E	I	Q	Q	A	I	K	K	A	K	D	D	V	I	N	V	I	Q	K	A	H	N	M	E	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	R	I	L	N	D	A	R	D	K	T	G	G	740
	<i>D. discoideum</i>	686	A	K	V	T	L	T	I	S	S	A	K	N	Q	V	K	E	L	I	I	K	A	Q	N	K	Q	F	E	C	Q	P	G	K	S	V	I	E	T	F	E	Q	K	V	N	Q	V	L	N	K	A	R	D	T	A	G	S	740
	<i>C. elegans</i>	694	L	D	I	Q	N	T	I	R	K	A	K	D	D	V	D	V	I	E	K	A	H	N	D	L	E	P	T	P	G	N	T	L	R	Q	T	F	E	N	Q	V	N	Q	I	L	N	D	A	R	D	R	T	G	S	748		
	<i>S. cerevisiae</i>	677	R	E	I	T	E	T	I	E	A	E	A	K	K	V	L	D	V	T	K	E	A	Q	N	D	L	T	A	K	H	G	M	T	L	R	S	E	F	D	N	V	V	R	F	L	N	E	A	R	D	K	A	G	R	S	731	
	<i>S. pombe</i>	683	K	E	V	T	R	T	V	K	E	A	R	R	Q	V	A	E	C	I	Q	D	A	Q	H	N	R	L	K	P	E	P	G	M	T	R	E	S	F	E	A	K	V	S	R	I	L	N	Q	A	R	D	N	A	G	R	S	737

Conservation of RPB1 funnel helix residues that interact with NELF-C

[illegible]**b**

Conservation of NELF-C residues that interact with RPB1

		485	488	494									524	526	531												
<i>H. sapiens</i>	480	EHSQ	LDVM	-EQLE	LKK	TLLDR	RMVH	LLSR	GYVL	LPVVS	YIRK	CLEK	LD	TD	ISL	IRYF	TEVLD	539									
<i>M. musculus</i>	481	EHSQ	LDVM	-EQLE	LKK	TLLDR	RMVH	LLSR	GYVL	LPVVS	YIRK	CLEK	LD	TD	ISL	IRYF	TEVLD	540									
<i>X. laevis</i>	472	EHSQ	LDVM	-EQLE	LKK	TLLDR	RMVH	LLSR	GYVL	LPVVT	YIRK	CLEK	LD	TD	ISL	IRYF	TEVLD	531									
<i>D. rerio</i>	469	EHSQ	LDVM	-EQLE	LKK	TLLDR	RMVH	LLSR	GYVL	LPVVG	YIRK	CLEK	LD	TD	ISL	IRYF	TEVLD	528									
<i>D. melanogaster</i>	469	KQDE	LEIL	-VQLEM	KK	MLDR	RMVH	NLLTR	GC	GV	PVLR	YIKQ	CA	ED	TD	ISL	IRYF	TEVLE	528								
<i>D. discolorum</i>	521	EAQD	LES	L	AILE	MRK	NV	IDN	IV	YLF	SC	GV	Y	PL	LD	TI	ES	WA	P	-KID	PSL	TRY	F	IN	QV	LD	578

C

NELF-C residues
contacting trigger loop

RPB1 Trigger
loop tip

<i>H. sapiens</i>	477	F	E	T	E	H	S	Q	L	D	V	M	E	Q	L	E	L	492	<i>H. sapiens</i>	1102	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1117
<i>S. scrofa</i>	477	F	E	T	E	H	S	Q	L	D	V	M	E	Q	L	E	L	492	<i>S. scrofa</i>	1098	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1113
<i>M. musculus</i>	478	F	E	T	E	H	S	Q	L	D	V	M	E	Q	L	E	L	493	<i>M. musculus</i>	1102	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1117
<i>X. laevis</i>	469	F	E	T	E	H	S	Q	L	D	V	M	E	Q	L	E	L	484	<i>X. laevis</i>	1100	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1115
<i>D. rerio</i>	466	F	E	T	E	H	S	Q	L	D	V	M	E	Q	M	E	L	481	<i>D. rerio</i>	1099	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1114
<i>D. melanogaster</i>	466	F	E	S	K	Q	D	E	L	E	I	L	V	Q	L	E	M	481	<i>D. melanogaster</i>	1094	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1109
<i>D. discoideum</i>	518	F	E	V	E	A	Q	D	L	E	S	L	A	I	L	E	M	533	<i>D. discoideum</i>	1098	M	T	L	N	T	F	H	Y	A	G	V	S	A	K	N	V	1111

d

d Lack of conservation of human RPB1 NELF associating residues among human Pol I (RPA1) and Pol III (RPC1) large subunits

Each of conservation of human RPA1 and RPA2, accelerating residues among human RPA1 (RPA1) and RPA2 (RPA2) large subunits

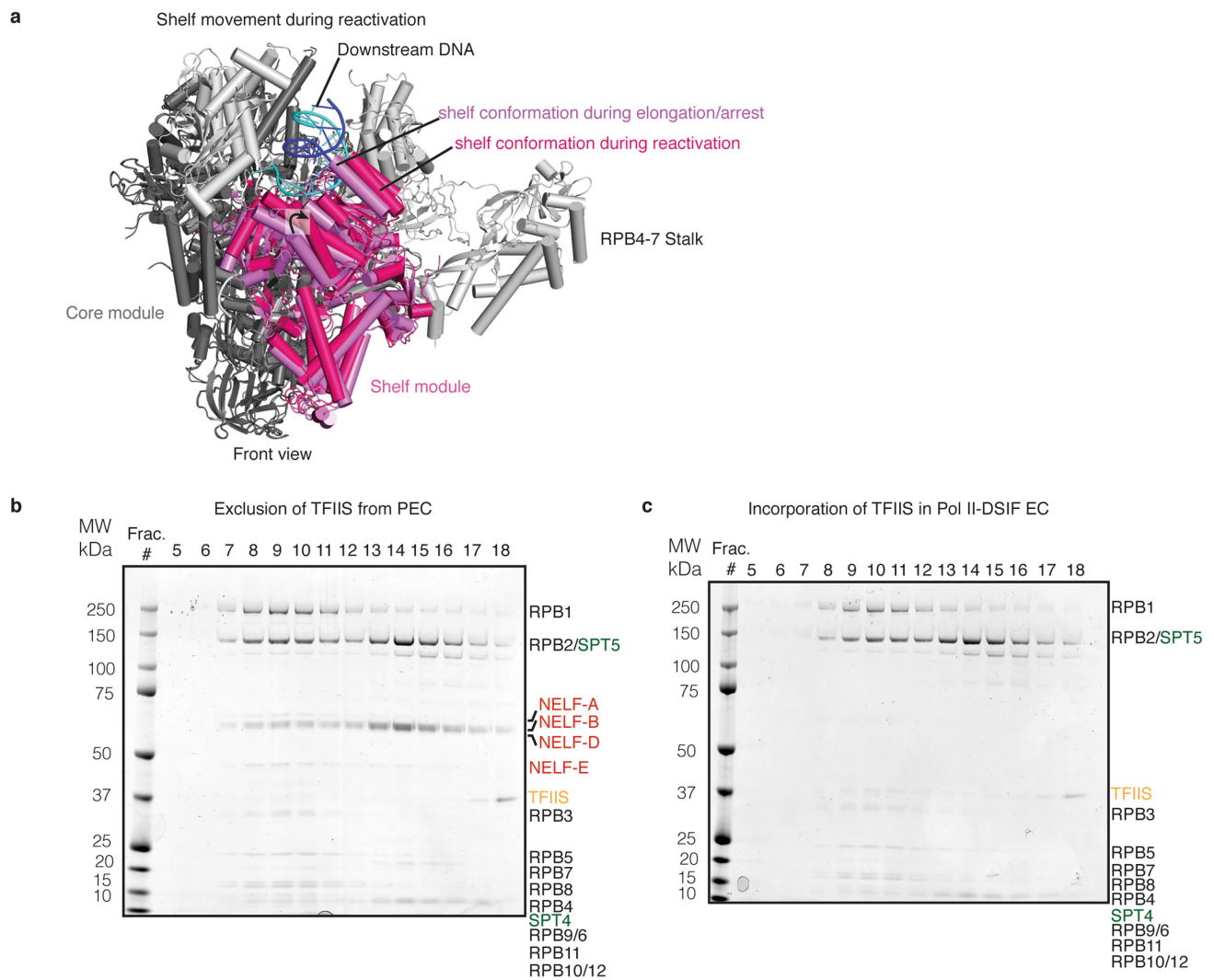
RPA1 803 LVKPKADVKRQRIIEETHCGPQAVRAALNLPEAASYDEVRGKWDAAHLGKGDQDFNMI DLK FKEEVNHY SNEI LNK A879
RPA2 693 IADSKTYQDIQNTIK-----KAKQDVI EVI EKAHNNE L-----EPT PGNT LRQT FENQVNRILNDARDK 751
RPA1 704 TPGQGLLKAYEILN-----AGYKKCDEYI LEALNTGKL-----QQQGCTAEET LEALI LKELSVI RDH762

1149 1152

RPA1 1261 TPMMSPVPL-NTKKALKRVKS 1280
RPA2 1136 TPLSLTVFL LGQSARDAERAKD 1156
RPA1 1082 TPIITTAOL--DKDDADYARL 1100

Extended Data Fig. 8 | Conservation of Pol II and NELF elements. Sequence alignments were made using MAFFT⁷⁶ and were visualized in Jalview⁷⁷. Sequences elements are coloured by identity. Darker shades of blue indicate higher levels of identity. Red boxes demarcate the interacting residue. **a**, Conservation of RPB1 funnel helix and shelf module residues that interact with NELF-C. Organisms that encode for NELF are indicated.

b, Conservation of NELF-C residues that interact with RPB1 funnel helix and shelf module residues. **c**, Conservation of NELF-C residues that interact with the RPB1 trigger loop. **d**, Conservation of Pol I (RPA1), Pol II (RPB1) and Pol III (RPC1) large subunits and putative NELF-C interaction interface.



Extended Data Fig. 9 | TFIIS does not interact with the PEC. **a**, Shelf movement relative to the Pol II core during reactivation. An arrested Pol II crystal structure (PDB ID: 3PO2) and the crystal structure of its reactivation intermediate (PDB ID: 3PO3) were aligned on their Pol II core modules^{25,31} (dark grey). The shelf module (pink) rotates away from the core module during reactivation. **b**, TFIIS does not bind the PEC. Fractions from size-exclusion chromatography with Pol II, DSIF, NELF

and TFIIS. The EC was incubated with DSIF, NELF and TFIIS and applied to a Superose 6 column. The PEC is formed, but TFIIS does not migrate with the PEC. The experiment was performed twice. **c**, TFIIS binds the Pol II-DSIF EC. Fractions from size-exclusion chromatography with Pol II, DSIF and TFIIS. The EC was incubated with DSIF and TFIIS. A stable Pol II-DSIF-TFIIS EC is formed. The experiment was performed twice.

Extended Data Table 1 | Components of the PEC

a.

Component	Subunit	Construct residues (aa) / scaffold length (nt)	Mass (kDa)	UniProt/Genbank identifier
	RPB1	1-1970	217.2	XP_020923484.1
	RPB2	1-1174	133.8	XP_003129085.4
	RPB3	1-275	31.4	XP_003355849.1
	RPB4	1-142	16.3	XP_020932152.1
	RPB5	1-210	24.6	XP_003354010.1
<i>S. scrofa</i> Pol II	RPB6	1-127	14.4	XP_003481589.1
	RPB7	1-172	19.2	XP_013849657.1
	RPB8	1-150	17.1	NP_001230270.1
	RPB9	1-125	14.5	NP_001192333.1
	RPB10	1-67	7.6	XP_003122432.1
	RPB11	1-117	13.2	XP_003124442.2
	RPB12	1-58	7.0	XP_003355060.1
	NELF-A	1-528	57.3	Q9H3P2-1
	NELF-B	1-580	65.7	Q8WX92-1
NELF	NELF-D ^a	1-581	65.5	Q8IXH7-4
	NELF-E	1-380	43.2	P18615-1
	SPT4 ^a	1-117	13.2	P63272-1
DSIF	SPT5	1-1087	121.0	O00267-1
	Template	48	14.9	
Nucleic acid	Non-template	48	14.7	
	RNA ^b	46	15.3	
Final	18 polypeptides, 3 nucleic acids	7860 aa, 142 nt	927.1	

b.

Amino acid substitutions between pig and human Pol II

Subunit	Position	Pig residue	Human residue
RPB2	882	Gly	Ser
RPB3	75	Thr	Ile
RPB3	140	Ser	Asn
RPB6	126	Ser	Thr

a. List of all protein and nucleic acid components of the PEC. For details of the complex assembly and composition, refer to the main text and Methods. aa, amino acids; nt, nucleotides; kDa, kilodalton. ^aConstruct possesses three or four residual amino acids from the TEV or 3C protease cleavage site, respectively, that are not reported in this table. ^bBears 5'-6 FAM label, and the mass of label is included in the molecular weight. **b.** Amino acid substitutions between human and *S. scrofa* Pol II.

Extended Data Table 2 | Cryo-EM data collection, refinement and validation statistics

	Map 1 (EMDB-0038) (PDB 6GML)	Map 2 (EMDB-0039)	Map 3 (EMDB-0040)	Map 4 (EMDB-0041)	Map 5 (EMDB-0042)
Data collection and processing					
Magnification	165,000	165,000	165,000	165,000	165,000
Voltage (kV)	300	300	300	300	300
Electron exposure (e ⁻ /Å ²)	34-47	34-47	34-47	34-47	34-47
Defocus range (μm)	0.25-4	0.25-4	0.25-4	0.25-4	0.25-4
Pixel size (Å)	1.2277 (binned from 0.81)	1.2277 (binned from 0.81)	1.2277 (binned from 0.81)	1.2277 (binned from 0.81)	1.2277 (binned from 0.81)
Symmetry imposed	C1	C1	C1	C1	C1
Initial particle images (no.)	2,347,915	2,347,915	2,347,915	2,347,915	2,347,915
Final particle images (no.)	162,269	73,131	40,960	29,197	126,776
Map resolution (Å)	3.3	3.4	3.4	3.9	3.3
FSC threshold	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	3-8.2	3.1-9	3.1-9.1	3.4-10.8	3.0-7.4
Refinement					
Initial model used (PDB code)	5OIK, 5L3X				
Model resolution (Å)	3.7				
FSC threshold	0.5				
Model resolution range (Å)	3-8.2				
Map sharpening <i>B</i> factor (Å ²)	-65				
Model composition					
Non-hydrogen atoms	44785				
Protein residues	5623				
Ligands	10				
<i>B</i> factors (Å ²)					
Protein	40.77				
Ligand	72.24				
R.m.s. deviations					
Bond lengths (Å)	0.008				
Bond angles (°)	1.254				
Validation					
MolProbity score	1.79				
Clashscore	7.10				
Poor rotamers (%)	0.50				
Ramachandran plot					
Favored (%)	94.10				
Allowed (%)	5.86				
Disallowed (%)	0.04				

Extended Data Table 3 | RNA Pol II–NELF interactions

NELF subunit	Domain	Residue range	Pol II subunit	Domain	Residue range
NELF-A	NELF-AC dimer	D23	RPB8	β -barrel	D23
NELF-B	Staircase		RPB5	Jaw	D67
NELF-C	NELF-AC dimer	E400	RPB1	Funnel	K708
	NELF-AC dimer	K402		Funnel	D712
	NELF-AC dimer	R439		Funnel	D747
	NELF-AC dimer	D485		Funnel	R743
	NELF-AC dimer	D488		Funnel	R743
	NELF-AC dimer	K494		Jaw	E1152
	NELF-AC dimer	D524		Jaw	R1149
	NELF-AC dimer	D526		Jaw	R1149
	NELF-AC dimer	D531		Jaw	E1152
	NELF-AC dimer	D560		Jaw	R1149

Selected interacting residues in Pol II and NELF and their respective domains are indicated. Note the highly polar nature of the Pol II–NELF interface.

Structure of activated transcription complex Pol II–DSIF–PAF–SPT6

Seychelle M. Vos¹, Lucas Farnung¹, Marc Boehning¹, Christoph Wigge¹, Andreas Linden^{2,3}, Henning Urlaub^{2,3} & Patrick Cramer^{1*}

Gene regulation involves activation of RNA polymerase II (Pol II) that is paused and bound by the protein complexes DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF). Here we show that formation of an activated Pol II elongation complex in vitro requires the kinase function of the positive transcription elongation factor b (P-TEFb) and the elongation factors PAF1 complex (PAF) and SPT6. The cryo-EM structure of an activated elongation complex of *Sus scrofa* Pol II and *Homo sapiens* DSIF, PAF and SPT6 was determined at 3.1 Å resolution and compared to the structure of the paused elongation complex formed by Pol II, DSIF and NELF. PAF displaces NELF from the Pol II funnel for pause release. P-TEFb phosphorylates the Pol II linker to the C-terminal domain. SPT6 binds to the phosphorylated C-terminal-domain linker and opens the RNA clamp formed by DSIF. These results provide the molecular basis for Pol II pause release and elongation activation.

Transcription of metazoan protein-coding genes is regulated during the elongation phase by promoter-proximal pausing and subsequent release of paused Pol II into elongation¹. During pausing, the Pol II elongation complex binds DSIF, which is composed of subunits SPT4 and SPT5, and four-subunit NELF^{2,3}. The recently solved structure of the mammalian Pol II–DSIF elongation complex showed that DSIF forms clamps around upstream DNA and exiting RNA⁴. Similar results were obtained for a related yeast complex⁵. In our accompanying paper⁶ we report the structure of the paused Pol II–DSIF–NELF elongation complex (PEC), which shows that NELF binds the polymerase funnel and the open trigger loop, an element of the Pol II active site that closes to stimulate nucleotide addition. The PEC structure adopts an inactive state with a tilted DNA–RNA hybrid that impairs binding of the next nucleoside triphosphate (NTP) substrate. These results suggest possible mechanisms for NELF-stabilized Pol II pausing, but the molecular basis of pause release and formation of an activated elongation complex are not known.

The release of paused Pol II into elongation requires the positive transcription elongation factor b (P-TEFb), which comprises the kinase CDK9 and the predominant cyclin, T1^{7,8}. P-TEFb phosphorylates DSIF, NELF and the C-terminal domain (CTD) of the large Pol II subunit RPB1^{9–13}. The elongation factor PAF1 complex (PAF) was recently implicated in Pol II pausing and pause release^{14,15}, although its role is not clear. PAF contains the subunits PAF1, LEO1, CTR9, CDC73 and WDR61^{16–18} and is required for transcription elongation through chromatin^{19,20}. SPT6 is another conserved elongation factor that is also required for chromatin transcription^{21,22}. SPT6 stimulates transcription elongation in vitro²³ and in vivo²¹.

Here we demonstrate that P-TEFb kinase activity enables the formation of a stable complex of the Pol II–DSIF elongation complex with PAF and SPT6 in vitro. We determined the cryo-EM structure of the resulting 20-subunit activated Pol II–DSIF–PAF–SPT6 elongation complex, which we call EC*. Comparison of the EC* structure with the accompanying structure⁶ of the PEC elucidates how NELF is displaced for Pol II release from pausing, and how Pol II is activated for productive RNA elongation and chromatin passage.

Formation of activated elongation complex EC*

In the accompanying paper⁶, we used an RNA extension assay to recapitulate the stabilising function of NELF in the pausing of a Pol II–DSIF elongation complex in vitro⁶. To understand how Pol II is released from pause sites, we extended this assay and additionally purified recombinant human P-TEFb, a catalytically inactive P-TEFb mutant (CDK9(D149N), hereafter referred to as P-TEFb(D149N)), PAF and SPT6 (Methods, Extended Data Fig. 1a–e, Extended Data Table 1). Elongation complexes were formed on a DNA–RNA scaffold (hereafter denoted the modified pause scaffold, Extended Data Fig. 1b) that enabled Pol II pausing after the addition of CTP and GTP, and enabled the use of ATP solely as a kinase substrate. Incubation of the PEC with active P-TEFb and ATP (Extended Data Fig. 1f, g) had no effect on pausing, as previously observed²⁴. However, when PAF was also included, RNA extension beyond the pause site was facilitated (Extended Data Fig. 1h, i). When both PAF and SPT6 were included, RNA extension was strongly stimulated (Extended Data Fig. 1j, k). These results show that PAF can reverse NELF-stabilized Pol II pausing in vitro when active P-TEFb and ATP are present, and that elongation is further stimulated when SPT6 is also present.

We then carried out RNA extension assays in the absence of NELF. When DSIF, PAF, SPT6, P-TEFb and ATP were added to the Pol II elongation complex, RNA extension was stimulated in a time- and concentration-dependent manner (Fig. 1a, Extended Data Fig. 2a–c). Stimulation was not observed when the mutant P-TEFb(D149N) was used (Fig. 1a, Extended Data Fig. 2a–c). When DSIF was absent, PAF and SPT6 stimulated elongation only modestly when incubated with active P-TEFb and ATP, whereas incubation with PAF or SPT6 alone had no effect (Extended Data Fig. 2d, e). This indicates that there is a functional interaction between PAF and SPT6, which is consistent with previous observations^{25–27}. These results demonstrate that stimulated RNA extension in vitro requires the presence of DSIF, PAF, SPT6, active P-TEFb and ATP.

On the basis of these functional results, we used size-exclusion chromatography to assess whether we could form a stable, activated Pol II elongation complex in vitro (Fig. 1b, Methods, Extended Data Fig. 2f–j). To enable subsequent structure determination, we used a nucleic acid

¹Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Göttingen, Germany. ²Max Planck Institute for Biophysical Chemistry, Bioanalytical Mass Spectrometry, Göttingen, Germany. ³University Medical Center Göttingen, Institute of Clinical Chemistry, Bioanalytics Group, Göttingen, Germany. *e-mail: patrick.cramer@mpibpc.mpg.de

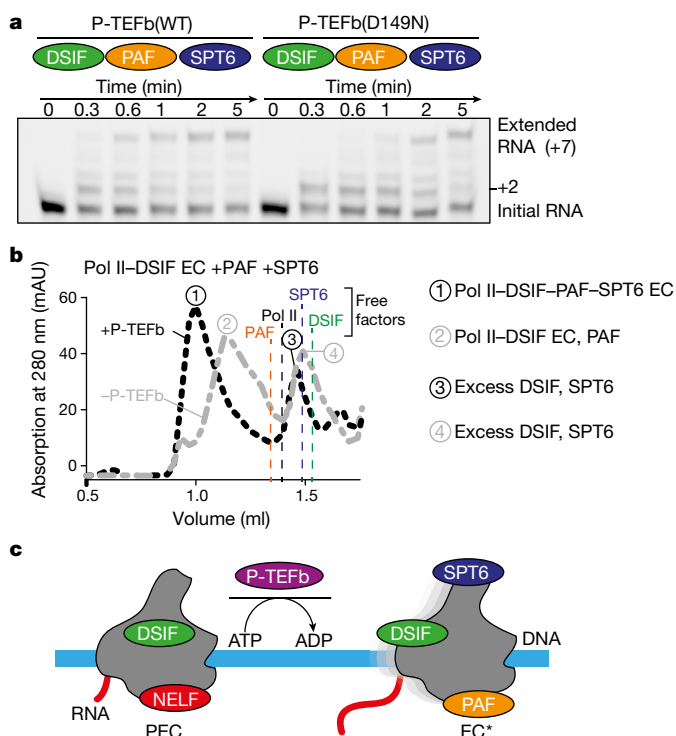


Fig. 1 | Formation of the EC* requires P-TEFb kinase. **a**, DSIF, PAF and SPT6 (75 nM) were incubated with Pol II (75 nM) on the modified pause scaffold (50 nM) (Extended Data Fig. 1b). Wild-type (WT) P-TEFb (left) or the inactive mutant P-TEFb(D149N) (right) (100 nM) and 1 mM ATP were added 15 min before transcription was initiated by the addition of 10 μ M GTP and CTP. Reactions were quenched at various times as indicated. Experiments were performed three times. **b**, Formation of the EC*. The Pol II–DSIF elongation complex assembled on the EC* scaffold was incubated with PAF and SPT6, either in the presence (+) or in the absence (–) of P-TEFb and ATP, and the resulting complexes were separated by size-exclusion chromatography (dashed lines). Dashed vertical lines mark the elution peaks of the free factors. Experiments were performed three times. **c**, Schematic showing conversion of the paused Pol II–DSIF–NELF elongation complex (PEC) to the activated Pol II–DSIF–PAF–SPT6 elongation complex (EC*).

scaffold that contained a DNA mismatch bubble (termed the EC* scaffold, Extended Data Fig. 1c). We found that DSIF readily bound the resulting elongation complex, whereas PAF and SPT6 association required P-TEFb and ATP (Fig. 1b, Extended Data Fig. 2f–j). These biochemical investigations led to the formation of the stable, activated elongation complex EC* (Fig. 1c) that contains Pol II, DSIF, PAF and SPT6, and elongates RNA efficiently.

Structure of the activated elongation complex EC*

After purification of EC* by size exclusion chromatography and mild crosslinking with glutaraldehyde, we determined its cryo-EM structure at a nominal resolution of 3.1 Å (Fig. 2, Supplementary Video 1, Extended Data Fig. 2j). 2D classification revealed densities on the Pol II surface (Extended Data Figs. 3, 4, Extended Data Table 2) and resulted in a 3D reconstruction from 374,964 particles. The core of Pol II was resolved at 2.6 Å. Elongation factors were resolved at lower resolutions (around 12 Å for the most flexible domains), and their corresponding densities were improved by focused classification and refinement (Extended Data Figs. 3–5, Methods). This led to a total of eight cryo-EM density maps that enabled us to fit available structures and homology models (Extended Data Fig. 3, Supplementary Table 1). Modelling was aided by lysine crosslinking data (Extended Data Fig. 6, Supplementary Tables 2–4). A total of 225 unique crosslinks were detected in structured regions, of which 210

fell into the permitted 30 Å range. The remaining 15 crosslinks were formed between mobile elements of the structure (Extended Data Fig. 6, Supplementary Table 2).

To complete the structure of the EC*, we determined the crystal structure of the isolated human SPT6 tandem SH2 (tSH2) domain at 1.8 Å resolution, and unambiguously docked this new structure into the corresponding density of EC* (Fig. 3, Methods, Extended Data Figs. 5f, 6f, 7, Extended Data Table 3). The resulting structure of EC* shows good stereochemistry and lacks only mobile regions, including the terminal regions of PAF1 and LEO1, most of CDC73, the acidic N-terminal region of SPT6, and the C-terminal extensions of SPT5, SPT6 and CTR9 (Supplementary Table 1).

PAF and SPT6 structure and contacts

DSIF, PAF and SPT6 are modular proteins that coat the outer surface of Pol II (Fig. 2). DSIF domains are arrayed around the Pol II cleft and the RNA exit tunnel⁴. PAF extends along the RPB2 side and docks on the Pol II funnel. PAF is anchored to the external domains of RPB2 by its PAF1–LEO1 dimerization module (Fig. 2b, c). The central PAF subunit CTR9 contains 19 tetratricopeptide repeats (TPRs; residues 41–750) that each form two antiparallel α -helices (Fig. 3a, Supplementary Table 6, Extended Data Fig. 5b). The CTR9 TPRs form a right-handed superhelix that extends from the Pol II subunit RPB11 along RPB8 via the polymerase funnel to the foot (Fig. 3a). The TPRs are followed by a pair of helices that create a ‘vertex’ and connect to a prominent ‘trestle’ helix in CTR9 (CTR9 residues 807–892) (Extended Data Fig. 5c). The trestle extends approximately 100 Å from the Pol II foot to subunit RPB5, where downstream DNA enters the Pol II cleft. The vertex and TPRs 13, 14 and 18 buttress the PAF subunit WDR61, which forms a seven-bladed β -propeller²⁸ and faces away from Pol II (Fig. 3a, Extended Data Figs. 5d, 8a). CDC73 is mobile except for an ‘anchor helix’ that binds CTR9 TPR 17 (Fig. 2d).

SPT6 binds the RPB4–RPB7 stalk on the RPB1 side of Pol II (Fig. 2c). The SPT6 core region is well resolved and resembles the structure of the yeast SPT6 core²⁹ (Fig. 3b, Extended Data Fig. 5g, Supplementary Table 7). Binding of the SPT6 core to the RPB4–RPB7 stalk includes an electrostatic interaction with the RPB7 β -strands C1–C3 (Extended Data Fig. 8b, c). These interactions of an elongation factor with RPB4–RPB7 bfit a role of the RPB4–RPB7 stalk not only during transcription initiation³⁰ but also during elongation^{31,32}. The SPT6 tSH2 domain is tethered flexibly to the SPT6 core and docks to Pol II at the site at which the CTD linker emerges to connect the CTD to the RPB1 body (CTD linker) (Fig. 2b).

Interactions are also observed between the elongation factors. The SPT5 domain KOWx–KOW4 contacts the SPT6 core, which explains the known SPT5–SPT6 genetic interaction³³ and a weak physical interaction between DSIF and SPT6^{23,34}. Low-pass filtering of the cryo-EM maps also revealed a density extending C-terminally from the SPT6 tSH2 domain to the CTR9 vertex and TPRs 18 and 19 (Extended Data Fig. 7e). This is consistent with known interactions between SPT6 and CTR9^{25–27}. Cryo-EM density and crosslinking data further indicate that the C-terminal tail of LEO1 contacts the upstream DNA and extends to the DNA clamp formed by the SPT5 NGN and KOW1 domains (Extended Data Figs. 5e, 6e), which explains previously reported PAF–DSIF interactions^{35–39}. Finally, initiation factors and elongation factors utilize similar regions of the Pol II surface for binding. The initiation factors TFIIB and TFIIE occupy similar regions to DSIF, whereas TFIIF and the coactivator complex Mediator engage regions bound by PAF and SPT6, respectively (Extended Data Fig. 8d). Taken together, these results show that DSIF, SPT6 and PAF are interconnected, coat a considerable portion of the Pol II surface, and could block the reassociation of initiation factors.

Release of NELF and paused Pol II

Comparison of the EC* structure with that of the PEC reported in the accompanying paper⁶ indicates that binding of NELF and PAF to Pol II is mutually exclusive (Fig. 2e, Supplementary Video 2). In particular,

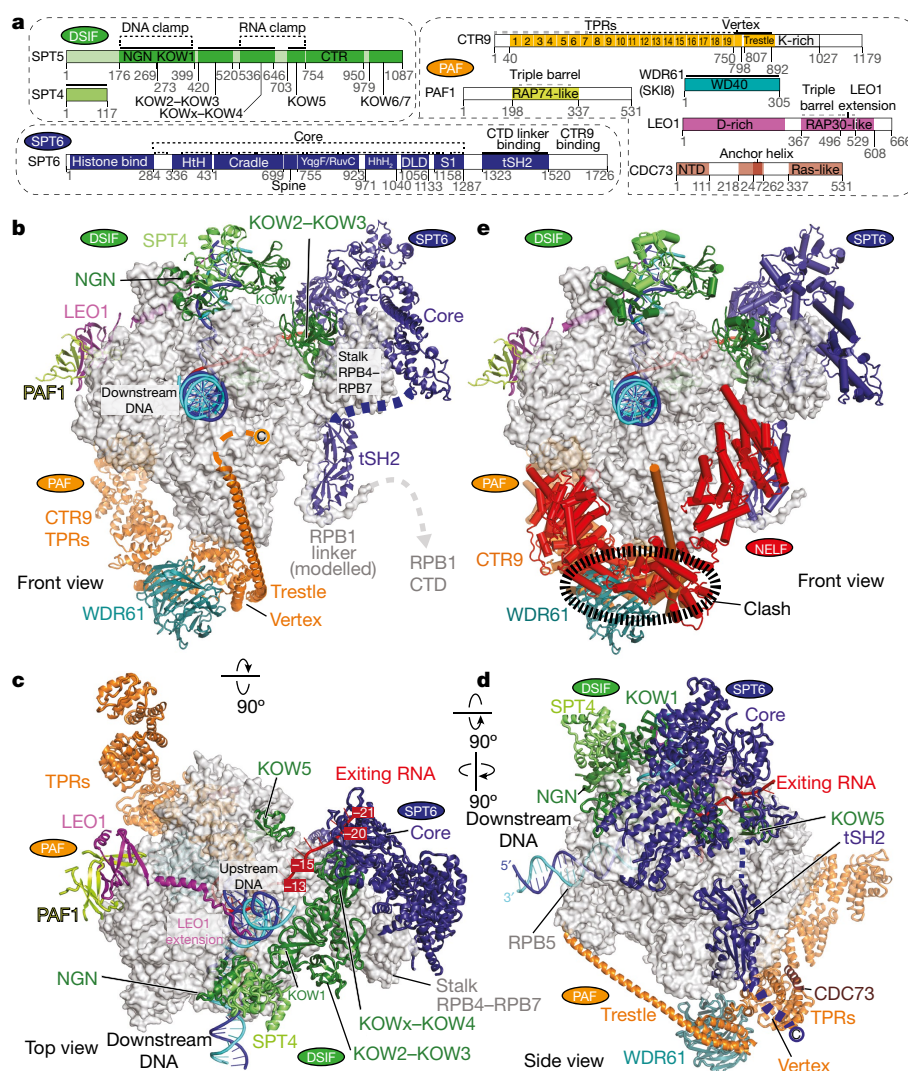


Fig. 2 | Cryo-EM structure of EC*. **a**, Domain architectures of DSIF, PAF and SPT6. The colour code is used throughout all figures. Black, dashed black, and grey lines indicate regions of the EC* structure that were included as atomic model, backbone model, or backbone model with unknown register, respectively. **b–d**, The EC* structure viewed from the Pol II front (**b**), side (**c**) and top (**d**). Pol II is shown as a silver surface;

DSIF, PAF and SPT6 are depicted as ribbon models. DNA template, DNA non-template and RNA are in blue, cyan and red, respectively. Dotted lines represent mobile protein regions. **e**, PAF and NELF binding sites overlap. The EC* and PEC structures were superimposed by aligning Pol II. The PAF subunits CTR9 (orange) and WDR61 (teal) clash with NELF (red).

NELF association with RPB8, the foot, and the protrusion is sterically incompatible with the binding of PAF to Pol II. To test whether NELF and PAF bind Pol II in a mutually exclusive manner, we incubated the Pol II elongation complex with DSIF, NELF, PAF, P-TEFb and ATP. Under these conditions, a stable Pol II–DSIF–PAF elongation complex was formed, and NELF was excluded (Extended Data Fig. 2k, l). Alternatively, when P-TEFb and ATP were omitted, a stable PEC was formed and PAF was excluded (Extended Data Fig. 2k, m). These data show that P-TEFb phosphorylation enables NELF release and PAF binding in our defined biochemical system. Together with our structural data, our results further indicate that PAF prevents the reassociation of NELF with the Pol II funnel.

Comparison of the PEC and EC* structures further shows a critical difference in the conformation of the DNA–RNA hybrid. The PEC adopts an inactive conformation with a tilted DNA–RNA hybrid that impairs the binding of the NTP substrate⁶, whereas the EC* adopts the active, post-translocated conformation with a free NTP-binding site (Fig. 3c). The trigger loop is observed in an open conformation in both the PEC and EC*. By contrast, NELF contacts the trigger loop, whereas PAF does not. The trigger loop in the EC* is therefore predicted to close easily after NTP binding, to stimulate the incorporation

of nucleotides and elongation of the RNA chain. These observations explain how NELF is displaced when Pol II is released from a pause site, and how the pause-stabilizing effects of NELF are overcome in EC*.

Changes in the DSIF DNA–RNA clamp

Our biochemical data show that PAF and SPT6 stimulate RNA extension (Fig. 1a, Extended Data Fig. 2a, d, e), consistent with published results of a stimulatory role of SPT6 *in vitro*²³ and *in vivo*²¹. Because neither PAF nor SPT6 reach the Pol II active site, the stimulatory effect is allosteric in nature. Comparison of the EC* structure with structures of the Pol II–DSIF elongation complex⁴ and the PEC⁶ revealed several conformational changes on the Pol II surface (Fig. 4a, Supplementary Video 3) that can explain the stimulatory effect of SPT6. The changes include repositioning of the RPB4–RPB7 stalk (Fig. 4b) and rearrangement of the SPT5 domains KOW2–KOW3 and KOWx–KOW4 (Fig. 4c). KOWx–KOW4 is rotated by 50° and moves away from exiting RNA by approximately 12 Å. This rearrangement breaks the previously observed contacts of the KOWx–KOW4 linker with RNA⁴, and thereby opens the RNA clamp of DSIF (Fig. 4c).

The rearrangement of KOW2–KOW3 and KOWx–KOW4 disrupts their previously observed interaction with KOW1 in the Pol II–DSIF

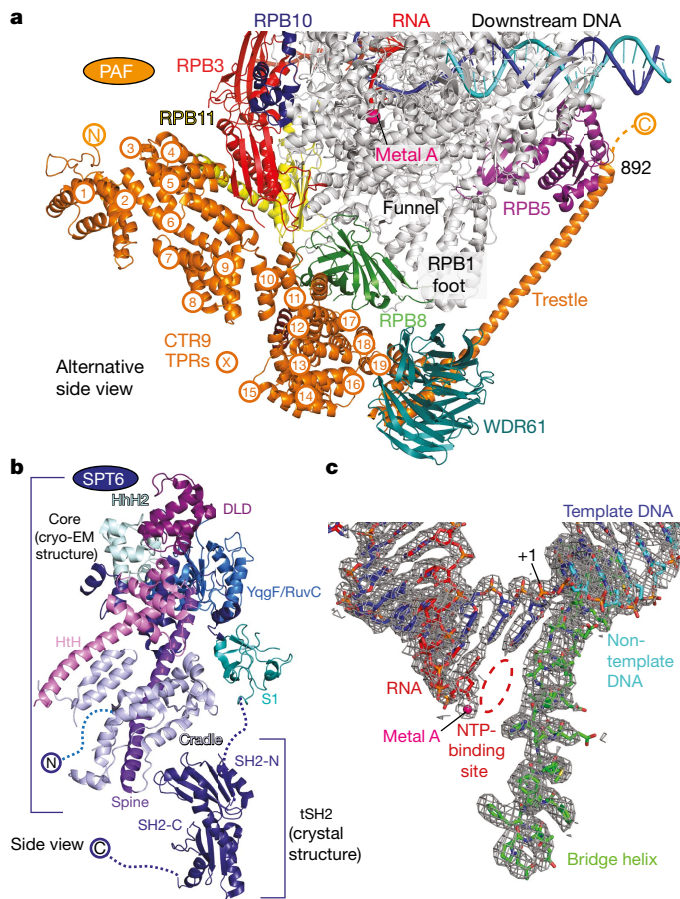


Fig. 3 | Structural details of EC*. **a**, PAF subunit CTR9 contacts Pol II and positions WDR61. Orange circles demarcate CTR9 TPR repeats. **b**, The structure of SPT6. Shown are the SPT6 core structure within EC* and the SPT6 tSH2 crystal structure. SPT6 domains are coloured in shades of blue. DLD, death-like domain; HhH₂, double-helix-hairpin-helix domain; HhH, helix-turn-helix domain. **c**, Nucleic acids in the EC* active site adopt a post-translocated state that can accept an incoming NTP substrate. Cryo-EM density from map A shown as mesh. The Pol II bridge helix and metal A are indicated.

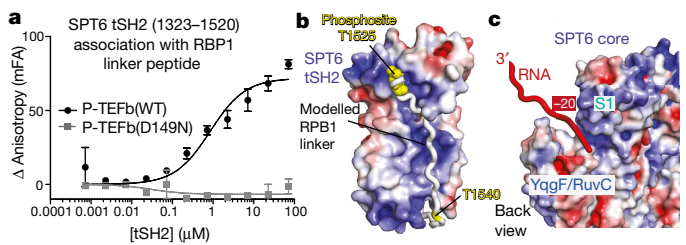


Fig. 5 | SPT6 binds CTD linker and RNA. **a**, Fluorescence anisotropy titration shows that the SPT6 tSH2 domain binds a CTD linker peptide that was incubated with P-TEFb and ATP ($K_{d,app} \sim 0.84 \mu\text{M} \pm 0.15$). Points represent the mean of three independent experiments and error bars are the standard deviation between the replicates. Source data, Supplementary Table 8. **b**, Model of the humanized yeast CTD linker (PDB ID: 5VKO⁴⁴) superimposed onto the human SPT6 tSH2 crystal structure (this work). The surface representation of the tSH2 domain is coloured according to charge (blue, positive; red, negative). The conserved RPB1 phosphorylation site Thr1525 is shown as a yellow sphere. The position of phosphorylated Thr1540 is indicated. **c**, Exiting RNA traverses a positively charged groove formed between SPT6 S1 and the YqgF/RuvC domains.

elongation complex⁴ and the PEC⁶, which results in a 40° rotation of KOW1 away from the upstream DNA. The rotation is accompanied by a movement of the upstream DNA, which is bent away from the protrusion (Fig. 4d). This generates a space between the upstream DNA and the protrusion that is occupied by the LEO1 C-terminal extension (Fig. 4e, Extended Data Fig. 5e). The path of the LEO1 extension is similar to that of a linker in the small subunit of the initiation factor TFIIF⁴⁰. LEO1 may thereby stabilize the upstream DNA and KOW1 in a new position that could facilitate the rewinding of upstream DNA. DNA rewinding is beneficial for elongation and is facilitated by the SPT5 homologue NusG in the bacterial system⁴¹. SPT4 and the SPT5 NGN domain remain fixed to keep the Pol II cleft closed and retain nucleic acids. Taken together, these findings show that PAF and SPT6 alter the DSIF DNA and RNA clamps, respectively, and stabilize the EC* conformation to stimulate elongation activity.

How P-TEFb triggers EC* formation

To investigate how P-TEFb triggers conversion of the PEC to EC*, we determined P-TEFb phosphorylation sites in vitro (Extended Data

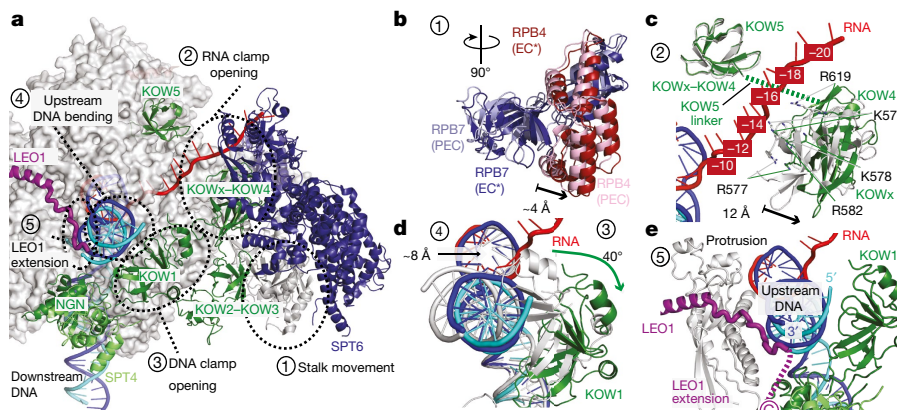


Fig. 4 | Conformational changes in DSIF. **a**, The structure of EC* viewed from the top. Regions with apparent conformational changes detected after superposition of the structures of EC* and the PEC⁶ are demarcated as dotted ovals and shown in detail in **b–e**. **b**, Movement of the RPB4–RPB7 stalk upon SPT6 binding. RPB4 is shown in red and RPB7 in blue, with corresponding pale shades depicting these subunits in the PEC. **c**, Opening of the DSIF RNA clamp. SPT5 domain KOWx–KOW4 is rotated away

from exiting RNA, whereas the position of KOW5 remains unchanged. **d**, Alteration of the DSIF DNA clamp. KOW1 is repositioned and upstream DNA is tilted in EC*. KOW1 rotates by 40° compared to its position in the PEC, whereas SPT4 and the SPT5 NGN domain remain in similar positions. **e**, The LEO1 C-terminal extension forms a wedge between the Pol II protrusion and the upstream DNA.

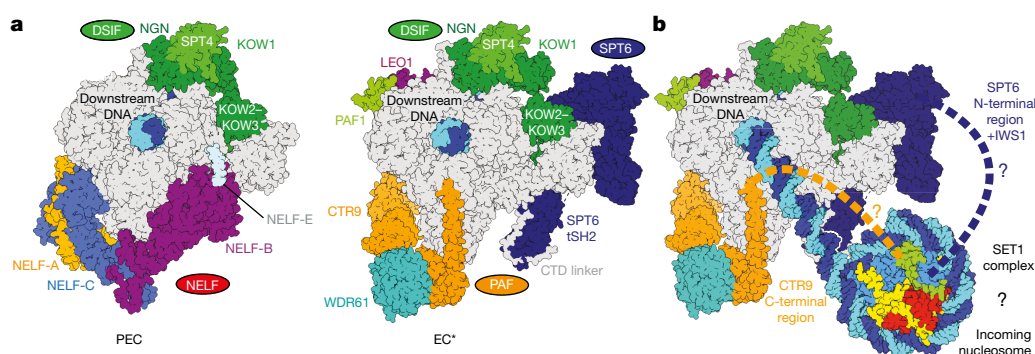


Fig. 6 | Comparison of the PEC and EC* structures. **a**, The structures of the PEC⁶ and the EC* are depicted schematically. NELF subunits are coloured orange (NELF-A), purple (NELF-B), blue (NELF-C) and light blue (NELF-E) as in the accompanying paper⁶. PEC nucleic acids, Pol II

and DSIF and EC* subunits are coloured as in Fig. 2a. **b**, Model of putative interactions between EC* and a downstream (incoming) nucleosome during chromatin transcription.

Fig. 9, Supplementary Table 5). We confirmed that P-TEFb phosphorylates the Pol II CTD and also mapped phosphorylation sites on DSIF, NELF, PAF and SPT6. We obtained 49 phosphorylation sites, of which ten per cent are known P-TEFb sites^{10,12,42}. Most of the phosphorylation sites are found in databases (Methods), which demonstrates that they are present in vivo. P-TEFb phosphorylates the NELF-A ‘tentacle’, which binds Pol II and is required for NELF-stabilized pausing^{6,43}. Phosphorylation of the NELF-A tentacle may facilitate NELF dissociation⁴². Phosphorylation of the SPT5 linker that connects KOWx–KOW4 and KOW5 may help to open the DSIF RNA clamp¹².

We next considered how P-TEFb enables the recruitment of SPT6. The SPT6 tSH2 domain lies adjacent to the CTD linker (Fig. 2b). It was recently reported that the yeast tSH2 domain binds to the phosphorylated CTD linker⁴⁴. We therefore tested whether the human CTD linker can be phosphorylated by P-TEFb. Indeed, P-TEFb could phosphorylate six human CTD linker residues in vitro (Extended Data Fig. 10a–f, Methods), of which Thr1525 corresponds to the yeast site Thr1471⁴⁴. Furthermore, a P-TEFb-treated CTD linker peptide bound the human tSH2 domain (Fig. 5a, Methods). We also found that the tSH2 domain is required for binding a linker-containing CTD variant, for SPT6 incorporation into EC*, and for the stimulation of elongation (Extended Data Fig. 10g–j), in accordance with previous work^{45,46}. Finally, modelling shows that the phosphorylated CTD linker can meander along a positively charged crevice of the human tSH2 domain (Fig. 5b, Extended Data Fig. 7f). These results show that P-TEFb phosphorylates the human CTD linker, and this enables SPT6 tSH2 binding and the stable docking of SPT6 to Pol II.

Binding of the CTD linker to SPT6 brings the CTD closer to the Pol II surface and the exiting RNA transcript. The exiting RNA in EC* passes through a positively charged groove formed between the S1 and the RuvC-like domains of the SPT6 core (Fig. 5c). Consistent with this structural observation, SPT6 modestly binds single-stranded nucleic acids (Extended Data Fig. 10k), and yeast SPT6 crosslinks to nascent RNA in cells⁴⁷. Factors involved in co-transcriptional RNA processing associate with the phosphorylated CTD, but also with the CTR of SPT5¹¹. The CTR is also phosphorylated by P-TEFb^{11,13} and extends from the KOW5 domain that lies adjacent to the exiting RNA. These findings reveal that the structural features involved in co-transcriptional RNA processing are clustered on the EC* surface (Extended Data Fig. 8e). The biochemical definition and structural characterization of EC* therefore provides a starting point for the analysis of elongation-coupled events such as co-transcriptional pre-mRNA processing.

Discussion

We report here that Pol II release from the paused state and elongation activation requires P-TEFb, PAF and SPT6 in vitro. We solved the structure of the activated Pol II–DSIF–PAF–SPT6 elongation complex, which we call EC*. Together with the accompanying paper⁶, our work provides the molecular basis for Pol II pausing, release of Pol II from the

paused state, and elongation activation. It also establishes a molecular framework for a detailed analysis of promoter-proximal transcriptional gene regulation by P-TEFb.

Comparison of the EC* structure with the PEC structure⁶ provides a model for understanding how paused Pol II is released into elongation and how elongation is activated (Fig. 6a). PAF sterically competes with NELF for binding to the Pol II funnel, and P-TEFb phosphorylation influences competition between NELF and PAF to facilitate NELF release and PAF binding. This is consistent with a requirement of PAF for pause release in cells¹⁴. P-TEFb phosphorylates not only the Pol II CTD, DSIF, NELF, PAF and SPT6, but also targets the CTD linker, to promote SPT6 binding. Interactions of PAF and SPT6 induce conformational changes in the DSIF clamps on upstream DNA and exiting RNA, respectively. These changes may promote DNA rewinding at the upstream edge of the transcription bubble, to drive the polymerase forward⁴⁸, and may facilitate RNA passage through the exit tunnel, to further stimulate elongation.

Finally, PAF and SPT6 have important roles in enabling transcription of the natural template, chromatin^{19,20}. Yeast PAF binds the major histone H3K4 methyltransferase Set1⁴⁹. PAF could reach downstream nucleosomes via its long trestle helix, and this may facilitate histone methylation by SET1 when Pol II approaches a nucleosome. SPT6 is a histone chaperone, and its N-terminal region binds to histones and the nucleosome-interacting protein IWS1^{45,46,50}. Although this SPT6 region is mobile in our structure, its location is constrained to the area between upstream and downstream DNA. This is consistent with the idea that SPT6 stores histones while Pol II transcribes through a nucleosome, thereby avoiding loss of histones and retaining epigenetic information during Pol II passage (Fig. 6b). Therefore, the EC* structure provides a starting point for analysing the mechanisms of chromatin transcription.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0440-4>.

Received: 19 March 2018; Accepted: 17 July 2018;
Published online 22 August 2018.

- Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
- Wada, T. et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* **12**, 343–356 (1998).
- Yamaguchi, Y. et al. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41–51 (1999).
- Bernecky, C., Plitzko, J. M. & Cramer, P. Structure of a transcribing RNA polymerase II–DSIF complex reveals a multidentate DNA–RNA clamp. *Nat. Struct. Mol. Biol.* **24**, 809–815 (2017).
- Ehara, H. et al. Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science* **357**, 921–924 (2017).

6. Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* <https://doi.org/10.1038/s41586-018-0442-2> (2018).
7. Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270**, 12335–12338 (1995).
8. Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451–462 (1998).
9. Marshall, N. F., Peng, J., Xie, Z. & Price, D. H. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J. Biol. Chem.* **271**, 27176–27183 (1996).
10. Fujinaga, K. et al. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol. Cell. Biol.* **24**, 787–795 (2004).
11. Yamada, T. et al. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol. Cell* **21**, 227–237 (2006).
12. Sansó, M. et al. P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. *Genes Dev.* **30**, 117–131 (2016).
13. Kim, J. B. & Sharp, P. A. Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. *J. Biol. Chem.* **276**, 12317–12323 (2001).
14. Yu, M. et al. RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science* **350**, 1383–1386 (2015).
15. Chen, F. X. et al. PAF1, a molecular regulator of promoter-proximal pausing by RNA polymerase II. *Cell* **162**, 1003–1015 (2015).
16. Zhu, B. et al. The human PAF complex coordinates transcription with events downstream of RNA synthesis. *Genes Dev.* **19**, 1668–1673 (2005).
17. Mueller, C. L. & Jaehning, J. A. Ctr9, Rtf1, and Leo1 are components of the Paf1/RNA polymerase II complex. *Mol. Cell. Biol.* **22**, 1971–1980 (2002).
18. Krogan, N. J. et al. RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell. Biol.* **22**, 6979–6992 (2002).
19. Kim, J., Guerma, M. & Roeder, R. G. The human PAF1 complex acts in chromatin transcription elongation both independently and cooperatively with SII/TFIIS. *Cell* **140**, 491–503 (2010).
20. Van Oss, S. B., Cucinotta, C. E. & Arndt, K. M. Emerging insights into the roles of the Paf1 complex in gene regulation. *Trends Biochem. Sci.* **42**, 788–798 (2017).
21. Ardehalii, M. B. et al. Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J.* **28**, 1067–1077 (2009).
22. Kaplan, C. D., Laprade, L. & Winston, F. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**, 1096–1099 (2003).
23. Endoh, M. et al. Human Spt6 stimulates transcription elongation by RNA polymerase II in vitro. *Mol. Cell. Biol.* **24**, 3324–3336 (2004).
24. Wada, T. et al. FACT relieves DSIF/NELF-mediated inhibition of transcriptional elongation and reveals functional differences between P-TEFb and TFIIF. *Mol. Cell* **5**, 1067–1072 (2000).
25. Kaplan, C. D., Holland, M. J. & Winston, F. Interaction between transcription elongation factors and mRNA 3'-end formation at the *Saccharomyces cerevisiae* GAL10–GAL7 locus. *J. Biol. Chem.* **280**, 913–922 (2005).
26. Adelman, K. et al. *Drosophila* Paf1 modulates chromatin structure at actively transcribed genes. *Mol. Cell. Biol.* **26**, 250–260 (2006).
27. Dronamraju, R. & Strahl, B. D. A feed forward circuit comprising Spt6, Ctk1 and PAF regulates Pol II CTD phosphorylation and transcription elongation. *Nucleic Acids Res.* **42**, 870–881 (2014).
28. Xu, C. & Min, J. Structure and function of WD40 domain proteins. *Protein Cell* **2**, 202–214 (2011).
29. Close, D. et al. Crystal structures of the *S. cerevisiae* Spt6 core and C-terminal tandem SH2 domain. *J. Mol. Biol.* **408**, 697–713 (2011).
30. Edwards, A. M., Kane, C. M., Young, R. A. & Kornberg, R. D. Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. *J. Biol. Chem.* **266**, 71–75 (1991).
31. Jasiak, A. J. et al. Genome-associated RNA polymerase II includes the dissociable Rpb4/7 subcomplex. *J. Biol. Chem.* **283**, 26423–26427 (2008).
32. Schulz, D., Pirkil, N., Lehmann, E. & Cramer, P. Rpb4 subunit functions mainly in mRNA synthesis by RNA polymerase II. *J. Biol. Chem.* **289**, 17446–17452 (2014).
33. Swanson, M. S. & Winston, F. SPT4, SPT5 and SPT6 interactions: effects on transcription and viability in *Saccharomyces cerevisiae*. *Genetics* **132**, 325–336 (1992).
34. Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev.* **12**, 357–369 (1998).
35. Qiu, H., Hu, C., Wong, C.-M. & Hinnebusch, A. G. The Spt4p subunit of yeast DSIF stimulates association of the Paf1 complex with elongating RNA polymerase II. *Mol. Cell. Biol.* **26**, 3135–3148 (2006).
36. Squazzo, S. L. et al. The Paf1 complex physically and functionally associates with transcription elongation factors in vivo. *EMBO J.* **21**, 1764–1774 (2002).
37. Chen, Y. et al. DSIF, the Paf1 complex, and Tat-SF1 have nonredundant, cooperative roles in RNA polymerase II elongation. *Genes Dev.* **23**, 2765–2777 (2009).
38. Liu, Y. et al. Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. *Mol. Cell. Biol.* **29**, 4852–4863 (2009).
39. Zhou, K., Kuo, W.-H. W., Fillingham, J. & Greenblatt, J. F. Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. *Proc. Natl Acad. Sci. USA* **106**, 6956–6961 (2009).
40. He, Y. et al. Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359–365 (2016).
41. Turtola, M. & Belogurov, G. A. NusG inhibits RNA polymerase backtracking by stabilizing the minimal transcription bubble. *eLife* **5**, e18096 (2016).
42. Lu, X. et al. Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II. *Nucleic Acids Res.* **44**, 6853–6867 (2016).
43. Narita, T. et al. Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Mol. Cell. Biol.* **23**, 1863–1873 (2003).
44. Sdano, M. A. et al. A novel SH2 recognition mechanism recruits Spt6 to the doubly phosphorylated RNA polymerase II linker at sites of transcription. *eLife* **6**, e28723 (2017).
45. Yoh, S. M., Cho, H., Pickle, L., Evans, R. M. & Jones, K. A. The Spt6 SH2 domain binds Ser2-P RNAPII to direct lws1-dependent mRNA splicing and export. *Genes Dev.* **21**, 160–174 (2007).
46. Yoh, S. M., Lucas, J. S. & Jones, K. A. The lws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev.* **22**, 3422–3434 (2008).
47. Battaglia, S. et al. RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases. *eLife* **6**, e25637 (2017).
48. Kireeva, M. et al. RNA–DNA and DNA–DNA base-pairing at the upstream edge of the transcription bubble regulate translocation of RNA polymerase and transcription rate. *Nucleic Acids Res.* **46**, 5764–5775 (2018).
49. Ng, H. H., Robert, F., Young, R. A. & Struhl, K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell* **11**, 709–719 (2003).
50. Bortvin, A. & Winston, F. Evidence that Spt6p controls chromatin structure by a direct interaction with histones. *Science* **272**, 1473–1476 (1996).

Acknowledgements We thank A. Kühn and M. Raabe for identifying phosphorylation sites by mass spectrometry, E. Wolf for pig thymus, F. Fischer and U. Neef for maintaining insect cell stocks, C. Oberthür and G. Kokic for assistance with protein purification, X. Liu and M. Ochmann for help with cloning and crystal refinement, H. S. Hillen and Swiss Light Source PXII for help with crystallographic data collection, and M. Geyer for sharing wild-type P-TEFb expression plasmids. S.M.V. was supported by an EMBO Long-Term Fellowship (ALTF 745-2014). H.U. was supported by the Deutsche Forschungsgemeinschaft (DFG SFB860). P.C. was supported by the Advanced Grant TRANREGULON (grant agreement 693023) of the European Research Council, and the Volkswagen Foundation.

Reviewer information *Nature* thanks K. Adelman, S. Darst and R. Landick for their contribution to the peer review of this work.

Author contributions S.M.V. designed and conducted all experiments unless stated otherwise. L.F. established and conducted SPT6 preparation and crystallized the SPT6 tSH2 domain. M.B. determined linker phosphorylation sites by mass spectrometry. C.W. assisted in cryo-EM data collection. A.L. performed crosslinking–mass spectrometry, supervised by H.U. P.C. supervised the research. S.M.V. and P.C. wrote the manuscript with input from L.F., M.B. and H.U.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0440-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0440-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.C.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and protein expression. DSIF, NELF and TFIIS were cloned as described in the accompanying paper⁶. cDNA clones encoding full-length human PAF subunits CDC73, WDR61, PAF1, LEO1 and CTR9 were obtained from the Harvard Plasmid Repository and the Medical Research Council Protein Phosphorylation and Ubiquitylation Unit. cDNAs were used as PCR templates for insertion into a modified pFASTbac vector (438-A, Addgene, 55218) via ligation-independent cloning⁵¹. CTR9 was cloned with a C-terminal tobacco etch virus (TEV) protease cleavable 6× His tag. All subunits were incorporated into a single plasmid by successive rounds of ligation-independent cloning. A cDNA clone encoding full-length human SPT6 (1–1726) (Harvard Plasmid Repository) was used as a PCR template for insertion of SPT6 into the 438-C vector (Addgene, 55220), which bears an N-terminal His6-MBP-tag (MBP, maltose-binding protein) followed by a TEV protease site. SPT6 ΔtSH2 (1–1297) was cloned by round-the-horn site-directed mutagenesis. The SPT6 tSH2 (1323–520) was amplified from cDNA and cloned into the 438-C vector.

A baculovirus expression plasmid encoding P-TEFb (CDK9 1–372, CYCT1 1–272) (pACEBac1, Geneva Biotech) was a gift from M. Geyer (University of Bonn). CDK9 is tagged with an N-terminal His8 tag followed by a TEV protease cleavage site. CYCT1 is tagged N-terminally with a GST tag followed by a TEV protease cleavage site. The CDK9 mutation Asp149Asn was introduced by site-directed mutagenesis.

Bacmid, virus and protein production for PAF, SPT6 and P-TEFb were performed as previously described⁵². Sf9 (ThermoFisher), Sf21 (Expression Systems) and Hi5 (Expression Systems) cell lines were not tested for mycoplasma contamination and were not authenticated in-house. Hi5 cells expressing PAF or SPT6 or P-TEFb were collected by centrifugation, resuspended in lysis buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole, 1 mM dithiothreitol (DTT), 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine), flash-frozen and stored at −80 °C until purification.

Regions of human RPB1 corresponding to the linker and CTD (1488–1970, 1593–1970, 1488–1592) were amplified from an RPB1 cDNA clone and inserted into a modified pET24b vector with an N-terminal His6-MBP tag followed by a TEV site (Addgene, 29654, 1C vector) by ligation-independent cloning. The 1593–1970 variant was also cloned into a modified pGEX vector as previously described⁴⁷. RPB1 CTD variants were overexpressed in *Escherichia coli* BL21 (DE3) RIL cells grown in LB medium. Cells were grown at 37 °C and protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside when cells reached an optical density at 600 nm (OD₆₀₀) of around 0.5. Cells were collected after 3 h by centrifugation, resuspended in lysis buffer, flash-frozen and stored at −80 °C.

Protein purification. All steps were performed at 4 °C unless otherwise stated. Pol II, DSIF, NELF and TFIIS were purified as described in the accompanying paper⁶.

PAF was purified from 2–4 l of Hi5 expression. Cell pellets were lysed by sonication and cleared by centrifugation. Clarified lysate was filtered through 0.8 μm syringe filters and applied to a 5 ml HisTrap HP column (GE Healthcare Life Sciences) equilibrated in lysis 400 buffer (400 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine). The column was washed with ten column volumes of lysis 400 buffer, followed by three column volumes of lysis 800 buffer (800 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine) and three column volumes of lysis 400 buffer. The column was then equilibrated in low-salt buffer (150 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine). The HisTrap column was then removed, and the HiTrap Q was washed with five column volumes of low-salt buffer. The HiTrap Q column was then developed over a gradient into lysis 800 buffer. Peak fractions were assessed by SDS–PAGE and Coomassie staining. Fractions containing PAF were pooled and mixed with TEV protease⁵³ and lambda protein phosphatase. The protein was placed in a Slide-A-Lyzer 10 kDa molecular weight cutoff (MWCO) (ThermoFisher Scientific) and dialysed overnight against lysis buffer 400 containing 1 mM MnCl₂. The protein was then applied to a 5 ml HisTrap

column equilibrated in lysis buffer 400 to remove uncleaved protein, the His tag and TEV protease. The flow-through was collected and concentrated in 100 kDa MWCO Amicon Ultra Centrifugal Filters (Merck). The protein was then applied to a HiLoad S200 16/600 pg column equilibrated in SE buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol and 1 mM DTT). Peak fractions were assessed by SDS–PAGE and Coomassie staining. Pure peak fractions containing PAF were pooled and concentrated in a 100 kDa MWCO Amicon Ultra Centrifugal Filters (Merck), aliquoted, snap frozen and stored at −80 °C until use. The identity of individual subunits was confirmed by mass spectrometric analysis.

Wild-type SPT6, SPT6ΔtSH2 and the SPT6 tSH2 were purified from 1.2 l of Hi5 cells. Cell pellets were lysed by sonication and cleared by centrifugation. Clarified lysate was filtered through 0.8 μm syringe filters and applied to a 5 ml HisTrap HP column (GE Healthcare Life Sciences) equilibrated in lysis buffer. The column was washed with five column volumes of lysis buffer followed by two column volumes of high-salt 1000 buffer (1,000 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine) and two column volumes of lysis buffer. The HisTrap column was then attached to a 15 ml amylose column equilibrated in lysis buffer (New England Biolabs), packed in an XK column (GE Healthcare Life Sciences). Protein was eluted from the HisTrap column directly onto the amylose column in nickel elution buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 500 mM imidazole pH 8.0, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine). After five column volumes, the HisTrap column was removed and the amylose column was washed with two column volumes of high-salt 1000 buffer followed by two column volumes of lysis buffer. The protein was eluted from the amylose column in amylose elution buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 117 mM maltose, 1 mM DTT, 0.284 μg ml^{−1} leupeptin, 1.37 μg ml^{−1} pepstatin A, 0.17 mg ml^{−1} PMSF and 0.33 mg ml^{−1} benzamidine). Peak fractions were assessed on SDS–PAGE and Coomassie staining. Peak fractions corresponding to full-length SPT6 were pooled, mixed with TEV protease and lambda protein phosphatase, and dialysed against lysis buffer with 1 mM MnCl₂ overnight in a 10 kDa MWCO Slide-A-Lyzer. The protein was then applied to a HisTrap column equilibrated in lysis buffer to remove the uncleaved protein, the His6-MBP tag and TEV protease. The flow-through was collected, concentrated in a 100 kDa MWCO Amicon Ultra Centrifugal Filter (Merck), and applied to a HiLoad S200 16/600 pg column equilibrated in SE buffer. Peak fractions were assessed by SDS–PAGE and Coomassie staining. Pure peak fractions containing SPT6 were pooled and concentrated in a 100 kDa MWCO Amicon Ultra Centrifugal Filters (Merck), aliquoted, flash-frozen and stored at −80 °C until use. The tSH2 was purified in essentially the same way with the exception that the protein was not subjected to amylose purification.

Wild-type P-TEFb and D149N P-TEFb were purified from 4 l of Hi5 expression. Cell pellets were thawed, lysed by sonication and cleared by centrifugation. Clarified lysate was filtered through 0.8 μm syringe filters and applied to a 5 ml HisTrap HP column (GE Healthcare Life Sciences) equilibrated in lysis buffer. The column was washed with five column volumes of lysis buffer followed by two column volumes of high-salt 1000 buffer and two column volumes of lysis buffer. The column was then washed with five column volumes of low salt buffer and connected to a HiTrap S column (GE Healthcare Life Sciences) equilibrated in low salt buffer. The HisTrap was developed with a gradient of Nickel 150 elution buffer. The flow-through was collected and peak fractions were analysed by SDS–PAGE followed by Coomassie staining. Peak fractions containing P-TEFb were pooled and mixed with TEV protease. The protein was dialysed against lysis buffer overnight in a 10 kDa MWCO Slide-A-Lyzer. The protein was removed from the Slide-A-Lyzer and applied to a HisTrap column equilibrated in lysis buffer. The flow through was collected and concentrated in a 10 kDa MWCO Amicon Ultra Centrifugal Filters (Merck) and applied to a HiLoad S200 16/600 pg column equilibrated in SE buffer. Peak fractions were assessed by SDS–PAGE and Coomassie staining. Pure peak fractions containing P-TEFb were pooled and concentrated in a 10 kDa MWCO Amicon Ultra Centrifugal Filters (Merck) to a final concentration of 5–10 μM, aliquoted, snap frozen, and stored at −80 °C until use.

His6-MBP RPB1 constructs (1488–1592, 1488–1970 and 1593–1970) were purified using a similar scheme. Cell pellets were thawed, lysed by sonication, and cleared by centrifugation. Lysates were filtered through 0.8 μm syringe filters and applied to 5 ml HisTrap columns equilibrated in lysis buffer. The columns were washed with ten column volumes lysis buffer, two column volumes of high-salt 1000 buffer followed by two column volumes of lysis buffer. The proteins were eluted from the HisTrap column with Nickel elution buffer over a gradient of nine column volumes. For the 1488–1592 construct, peak fractions were pooled and concentrated in 10 kDa MWCO Amicon Ultra Centrifugal Filters (Merck) and applied to a HiLoad S75 16/600 column equilibrated in SE buffer. For the 1488–1970 and 1593–1970 constructs, the HisTrap column was attached to an Amylose column equilibrated in lysis buffer as the HisTrap column was eluted. The HisTrap

column was then removed and the amylose column was washed with two column volumes of high-salt 1000 buffer followed by two column volumes of lysis buffer. The protein was eluted from the amylose column with amylose elution buffer. Peak fractions were concentrated in 30 kDa MWCO Amicon Ultra Centrifugal Filters (Merck) and applied to a HiLoad S200 16/600 pg column equilibrated in SE buffer. Peak fractions eluting from the S75 and S200 columns were assessed by SDS–PAGE followed by Coomassie staining. The protein constructs were concentrated as above, flash-frozen in liquid nitrogen and stored at -80°C until use.

RNA extension assays. Transcription assays were performed with complementary DNA scaffolds that were designed to disfavor ATP misincorporation^{54–56}. All oligos were purchased from Integrated DNA Technologies, resuspended in water (100 μM), flash-frozen in liquid nitrogen and stored at -80°C . The sequences used for transcription assays are as follows: Modified pause scaffold: template DNA 5'-CCA CAG GAA GAA CAG AAA CAA CGG GCG GAA CTA TGC CGG ACG TAC TGA CCA-3', non-template DNA 5'-Biotin-TTT TTG GTC AGT ACG TCC GGC ATA GTT CCG CCC GTT GTT TCT GTT CTT CCT GTG G-3', RNA 5'-/6-FAM-UUU UUU GGC AUA GUU-3'; EC* transcription scaffold: template DNA GTT TCC CCC AGC TCC CAG CTC CCT GCT GGC TCC GAG TGG GTT CTG CCG CTC TCA ATG G, non-template DNA CCA TTG AGA GCG GCA GAA CCC ACT CGG AGC CAG CAG GGA GCT GGG AGC TGG GGG AAA C, RNA 5'-/6-FAM-UUA AGG AAU UAA GUC GUG CGU CUA AUA ACC GGA GAG GGA ACC CAC U-3'. The modified pause scaffold contains 13 nucleotides of upstream DNA, 28 nucleotides of downstream DNA, a nine-base pair (bp) DNA–RNA hybrid, and six nucleotides of exiting RNA bearing a 5' 6-FAM label (Extended Data Fig. 1). The modified pause scaffold was derived from the pause scaffold (bacterial)^{6,56} and was altered to disfavor ATP misincorporation during incubation steps with P-TEFb⁵⁴. The EC* transcription scaffold contains 15 nucleotides of upstream DNA, 34 nucleotides of downstream DNA, a nine-base-pair DNA–RNA hybrid, and 37 nucleotides of exiting RNA bearing a 5' 6-FAM label. The EC* transcription scaffold has same sequence as the EC* scaffold but a matched DNA bubble and an additional ten bases of DNA at the downstream edge of the DNA. RNA extension assays performed on the EC* transcription scaffold resemble the activity observed on the modified pause scaffold (Extended Data Fig. 1l, m).

RNA and template DNA were mixed in equimolar ratios and were annealed by incubating the nucleic acids at 95°C for 5 min and then decreasing the temperature by $1^{\circ}\text{C min}^{-1}$ steps to a final temperature of 30°C in a thermocycler in a buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl_2 and 10% (v/v) glycerol. All concentrations refer to the final concentrations used in the assay. *S. scrofa* Pol II (75 nM) and the RNA–template hybrid (50 nM) were incubated for 10 min at 30°C , shaking at 300 rpm. The non-template DNA (50 nM) was added and the reactions were incubated for another 10 min. The reactions were then diluted to achieve final assay conditions of 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl_2 , 4% (v/v) glycerol and 1 mM DTT and were again incubated for 10 min. Factors were diluted in protein dilution buffer (300 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol and 1 mM DTT) and added to Pol II elongation complexes as serial dilutions (0–750 nM) or at a concentration of 75 or 150 nM for time-course experiments. Wild-type P-TEFb or the inactive P-TEFb mutant D149N was added (100 nM) with 1 mM ATP and incubated with Pol II and the elongation factors for 15 min at 30°C . Transcription reactions were initiated by adding GTP and CTP (10 μM) to permit elongation to position +7 (modified pause scaffold) or GTP, CTP and UTP (10 μM) (EC* transcription scaffold). Reactions were quenched after 1–2 min (titration experiments) and after (0–5 min) for time-course experiments in 2x stop buffer (6.4 M urea, 50 mM EDTA pH 8.0, 1x TBE buffer). Samples were treated with 4 μg of proteinase K for 30 min (New England Biolabs) and were separated by denaturing gel electrophoresis (8 μl of sample applied to an 8 M urea, 1x TBE, 20% Bis-Tris acrylamide 19:1 gel run in 0.5x TBE buffer at 300 V for 90 min). Products were visualized using the 6-FAM label and a Typhoon 9500 FLA Imager (GE Healthcare Life Sciences).

Gel images were quantified using ImageJ version 1.48v⁵⁷. The integrated density of the elongated product was measured using a box size of 0.35×0.15 cm. All integrated density values were normalized by subtracting the background integrated density from each gel. Graphs were prepared in GraphPad Prism version 6. Each bar or point represents the mean intensity from two or three individual replicates. Error bars reflect the standard deviation between the replicates. Source data for gel quantification can be found in Supplementary Table 8.

We observe extension from a fraction of the input RNA molecules. We attribute this to inefficient assembly of the elongation complex on the perfectly complementary scaffolds used in our assays. It was previously shown that only 10–50% of yeast Pol II molecules successfully assemble on perfectly complementary scaffolds^{58–61} owing to non-template DNA displacement of the RNA primer⁵⁸. Others have resolved the problem of displaced RNA primer by incorporating radioactive NTPs or by immobilizing non-template DNA containing complexes on beads. We chose to perform RNA extension experiments in bulk with a fluorescently

labelled RNA to maintain consistent Pol II concentrations across experiments and reproducibility in time-course experiments.

Analytical gel filtration. Elongation complexes were formed on a bubble scaffold with the following nucleic acid sequence (EC* scaffold): template DNA 5'-GCT CCC AGC TCC CTG CTG GCT CCG AGT GGG TTC TGC CGC TCT CAA TGG-3', non-template DNA 5'-CCA TTG AGA GCG GCC CTT GTG TTC AGG AGC CAG CAG GGA GCT GGG AGC-3', and RNA 5'-/6-FAM - UUA AGG AAU UAA GUC GUG CGU CUA AUA ACC GGA GAG GGA ACC CAC U-3'. Pol II elongation complexes were formed as described for the transcription assays (25 pmol final Pol II, 50 pmol RNA–DNA template, 100 pmol non-template DNA). Elongation factors were added in 2–3 molar excess relative to Pol II in a final buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 3 mM MgCl_2 , 1 mM DTT, and 4% (v/v) glycerol. Reactions that included P-TEFb (18 pmol) were supplemented with 1 mM ATP pH 7.5. Reactions were incubated for 30 min at 30°C . Samples were applied to a Superose 6 increase 3.2/300 column equilibrated in complex buffer (100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 3 mM MgCl_2 and 1 mM DTT). Peak fractions were analysed by SDS–PAGE followed by Coomassie staining.

Sample preparation for cryo-EM. Samples for cryo-EM were prepared essentially as described for analytical gel filtration runs. Final protein amounts used for complex formation were 112 pmol Pol II, 168 pmol RNA–template DNA hybrid, 200 pmol non-template DNA, 224 pmol PAF, DSIF and SPT6, 38 pmol P-TEFb and 313 pmol TFIIS (when included). Peak fractions corresponding to the complex were individually crosslinked with 0.1% (v/v) glutaraldehyde for 10 min on ice. Reactions were quenched with 8 mM aspartate and 2 mM lysine and were dialysed against a buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 20 mM Tris-HCl pH 7.5, 1 mM DTT and 3 mM MgCl_2 , in 20 kDa MWCO Slide-A-Lyzer MINI Dialysis Units for 6 h at 4°C . Sample from the peak (150–175 nM) was applied to R2/2 gold grids and R2/1 carbon grids (Quantifoil). The grids were glow-discharged for 45 s before applying 2 μl of sample to each side of the grid (4 μl total). After incubation for 10 s and blotting for 8.5 s, the grid was vitrified by plunging it into liquid ethane with a Vitrobot Mark IV (FEI Company) operated at 4°C and 100% humidity.

Cryo-EM data collection and data processing. Three separate datasets were collected, two of which were collected in the presence of TFIIS. TFIIS was included because of its reported role in stabilizing the association of PAF with Pol II¹⁹. Here we describe the structure lacking TFIIS. Cryo-EM data was collected on a FEI Titan Krios II transmission electron microscope operated at 300 keV. A K2 summit direct detector (Gatan) with a GIF quantum energy filter (Gatan) was operated with a slit width of 20 eV. Automated data acquisition was done with FEI EPU software at a nominal magnification of 130,000 \times , corresponding to a pixel size of 1.049 Å per pixel. Image stacks of 40 frames were collected over 10 s in counting mode. The dose rate was $3.4\text{--}4.7\text{ e}^- \text{ per } \text{\AA}^2 \text{ per s}$ for a total dose of $34\text{--}47\text{ e}^- \text{ per } \text{\AA}^2$. A total of 20,198 image stacks were collected.

Frames were stacked and subsequently processed with MotionCorr2⁶². CTF correction was performed with Gctf⁶³. Image processing was performed with RELION 2.1^{64,65}. Particles were auto-picked using projections of an initial reconstruction of PAF and DSIF bound to Pol II (data not shown) yielding 1,775,917 particle images. Particles were extracted using a box size of 360^2 pixels, normalized, and screened using iterative rounds of reference-free 2D classification resulting in 1,675,585 particles. Particles from each of the three datasets were initially processed separately. An initial reconstruction of Pol II bound to PAF and DSIF (not shown) was low-pass filtered to 50 Å and used for hierarchical 3D-classification with and without image alignment and 3D-refinement. Classes showing density for TFIIS were omitted. The best resolved, non-TFIIS bound classes from each dataset were selected and combined resulting in 374,964 particles (dataset 1: 101,509 particles, dataset 2 (TFIIS): 58,720 particles, dataset 3 (TFIIS): 214,745 particles). The combined particles were subjected to 3D refinement using a 50 Å low-pass-filtered map from a previous 3D-refinement resulting in a reconstruction with a resolution of 3.10 Å (map A). Some domains were not well resolved in the reconstruction, so 3D classifications without image alignment with applied soft masks around the regions of interest were performed. Masks were generated in Chimera (version 1.10.2) and RELION 2.1 around the DSIF NGN (map B, 3.10 Å), KOW1 (map C, 3.49 Å), and KOWx–KOW4 (map D, 3.20 Å) domains, the SPT6 core (map E, 3.28 Å) and SPT6 tSH2 (map F, 3.49 Å), upstream DNA (map G, 3.10 Å), and CTR9 (map H, 3.34 Å)⁶⁶. Particles containing the desired density were subjected to global 3D refinement. To further improve densities, focused refinement was used for regions of CTR9, SPT6 and the RPB4–RPB7 stalk. Focused refinements were performed by continuing global refinements after the first iteration of local searches and applying a soft mask. Masks were generated for three regions of CTR9/WDR61 comprising the N-terminal region, middle region, and C-terminal region/WDR61 (map H). The C-terminal region/WDR61 focused refinement resulted in a final resolution of 3.59 Å with an applied *B*-factor of -145.94 \AA^2 . Focused refinement for RPB4–RPB7 was performed using map H resulting in a final resolution

of 3.63 Å with an applied *B*-factor of -140.45 Å^2 . Focus refinement was performed on the SPT6 core using the same mask used for classification (map E) resulting in a resolution of 4.44 Å with an applied *B*-factor of -171.70 Å^2 . Post-processing of refined models was performed using automatic *B*-factor determination in RELION and reported resolutions are based on the gold-standard Fourier shell correlation 0.143 criterion⁶⁷ (applied *B*-factors (Å²): map A: -98.65 , map B: -90.81 , map C: -90.57 , map D: -90.87 , map E: -88.56 , map F: -109.1 , map G: -94.30 , map H: -86.13). Local resolution estimates were determined using a sliding window of 30² voxels with a Fourier shell correlation cutoff of 0.3 on sharpened and non-*B*-factor sharpened maps as previously described⁶⁸.

Model building. The structure of EC* was solved by first placing the structure of a bovine elongation complex into map A in Chimera⁶⁶ (PDB ID: 5OIK)⁴. Adjustments were made to the protein sequence, DNA sequence, and positioning of the upstream DNA in Coot⁶⁹. The human RPB4–RPB7 crystal structure (PDB ID: 2C35)⁷⁰ was placed into a focused refined version of map H in Chimera.

Human DSIF from a previously solved cryo-EM structure (PDB ID: 5OIK)⁴ was divided into five regions for modelling, corresponding to the SPT5 NGN and SPT4, KOW1, KOW2–KOW3, KOWx–KOW4 and KOW5 and placed into globally refined maps. KOW2–KOW3 and KOW5 were placed in map A by rigid-body fitting in Chimera. The NGN domain and SPT4 were placed in map B by rigid-body fitting in Chimera. KOW1 and KOWx–KOW4 were placed into map C and map D, respectively, by rigid-body fitting in phenix.real_space_refine⁷¹. Densities for all five PAF subunits are observed. PAF was modelled using the known crystal structure for WDR61 (PDB ID: 3OW8) and homology modelling for the remaining subunits. A model for CTR9 1–798 was generated with Robetta⁷² using PDB ID 4BUJ⁷³ as a template. Secondary structure predictions from SABLE⁷⁴ and PSIPRED⁷⁵ were used to confirm the model. TPRs were identified and validated with TPRpred⁷⁶. CTR9 807–892 was built de novo in Coot in a focused refined version of map H. Crosslinking restraints and densities from bulky residues such as Arg and Tyr were used as sequence markers. The Robetta model of CTR9 was divided into four parts corresponding to residues 1–303, 303–677, 678–750 and 750–798 and fit into map H or focused refined versions of map H. WDR61 was placed into a focused refined version of map H and is shown as an atomic model. The orientation was determined from crosslinking data (Extended Data Fig. 6). CTR9 and WDR61 were flexibility fit into focused refined versions of map H using VMD and MDFF⁷⁷. The N terminus of CTR9 (1–300) is not well resolved, and is modelled as a backbone trace with unknown register. Clear helical densities are observed for residues 301–750 and are shown as backbone traces.

The predicted structural similarity between the triple barrel dimerization domain of TFIIF and PAF1/LEO1 was used to generate a homology model for PAF1/LEO1⁷⁸. PAF1 (188–341) and LEO1 (367–608) were separately threaded through the RAP74 and RAP30 subunits of human TFIIF (PDB ID: 5IYC)⁴⁰, respectively, using Phyre2 (99.05/99.81% confidence of threading). These regions were chosen owing to their predicted secondary structure similarity to TFIIF. The threaded model for LEO1 was truncated to residue 497. The threaded models for PAF1 and LEO1 were aligned on the TFIIF structure⁴⁰ and placed into the corresponding density in UCSF Chimera. Residues 498–529 of LEO1 were built de novo in Coot using crosslinking restraints and secondary structure predictions. PAF1 and LEO1 residues are modelled as backbone traces with an unknown register.

CDC73 is the least well-resolved subunit in our structure. Extensive crosslinking between CDC73 and CTR9 and noisy density near CTR9 suggests that CDC73 is highly mobile. We observe an additional helix immediately adjacent to CTR9 TPR 17 that cannot be assigned to CTR9. Crosslinking data and secondary structure predictions assigned this ‘anchor helix’ to CDC73 residues 249–262. This assignment is consistent with biochemical experiments that have collectively shown that a region of CDC73 corresponding to residues 200–337 is required for its association with PAF^{79–81}.

We compared our PAF structure with the published yeast Pol II–PAF–TFIIS structure⁷⁸. The general Pol II binding surfaces of yeast and human PAF are shared; however, there are several notable differences between the structures. First, there is a substantial difference in subunit composition between yeast and human PAF^{16–18,82}. Yeast PAF stably associates with Rtf1 and does not associate with Ski8 (WDR61 in human), whereas the opposite is true for human PAF. Second, the yeast Ctr9 construct used for cryo-EM was severely truncated and lacked the trestle helix. The trestle helix makes additional contacts with Pol II that may confer stability. Last, the yeast structure was solved as a ternary structure with Pol II and TFIIS, which may have rendered the complex more flexible. DSIF, SPT6 and P-TEFb phosphorylation greatly stabilize human PAF association with Pol II (data not shown). Together, these differences may have contributed to the higher flexibility of the yeast structure and differences in subunit assignment.

To generate a model for SPT6, the human sequence for SPT6 was threaded through a crystal structure of the *Saccharomyces cerevisiae* Spt6 core region (PDB ID: 3PSI)²⁹ with Phyre2⁸³. The model generated by Phyre2 was flexibility fit into a focused refined version of map E using VMD and MDFF⁷⁷. The model was

manually adjusted in Coot. Most domains of the central region are easily resolved with the exception of the death-like domain, which is more flexible than the rest of the complex (Extended Data Fig. 4, 5). We modelled three of a total of nine human-specific short insertions within the SPT6 core. The core was modelled as a backbone trace. The crystal structure of the human tSH2 was placed into map F using rigid-body fitting in phenix.real_space_refine. A loop corresponding to residues 1385–1395 was removed. The CTD linker was modelled using a previously solved crystal structure of yeast SPT6 tSH2 with bound CTD linker (PDB ID: 5VKO⁴⁴). The CTD linker was mutated to the corresponding human sequence with Phyre2 and fit onto the human tSH2 by matching the cores of the yeast and human tSH2 crystal structures in PyMOL (Schrödinger LLC, version 1.8.6.0).

The model was manually adjusted in Coot⁶⁹ and refined with phenix.real_space_refine against a locally filtered, non-sharpened version of map A. The final model has 95.07% of residues in most-favoured regions of the Ramachandran plot according to MolProbity⁸⁴. The structure has a MolProbity score of 1.64. Figures were generated in PyMOL (Schrödinger LLC, version 1.8.6.0) and UCSF Chimera (version 1.10.2). Surface charge was calculated with PDB2PQR⁸⁵ and visualized with APBS⁸⁶ in PyMOL (Schrödinger LLC, version 1.8.2.3).

Crosslinking–mass spectrometry. Samples for crosslinking and mass spectrometry analysis were essentially prepared as those used for cryo-EM. Fractions containing EC* were pooled and mixed with 2 mM of BS3 dissolved in complex buffer (No Weigh Format, ThermoFisher Scientific). The protein was incubated for 30 min at 30 °C. The crosslinking reaction was quenched by adding 100 mM Tris-HCl pH 7.5 and 20 mM ammonium bicarbonate (final concentrations). The reaction was incubated for 15 min further at 30 °C. The protein was precipitated with 300 mM sodium acetate pH 5.2 and four volumes of acetone and incubated overnight at -20 °C . The protein was pelleted by centrifugation, briefly dried, and resuspended in 4 M urea and 50 mM ammonium bicarbonate.

Crosslinked proteins were reduced with 10 mM DTT for one hour at room temperature. Alkylation was performed by adding iodoacetamide to a final concentration of 40 mM, and incubating for 30 min in the dark at room temperature. After dilution to 1 M urea with 50 mM ammonium bicarbonate (pH 8.0), the crosslinked protein complex was digested with trypsin in a 1:50 enzyme-to-protein ratio at 37 °C overnight. Peptides were acidified with trifluoroacetic acid (TFA) to a final concentration of 0.5% (v/v), desalted on MicroSpin columns (Harvard Apparatus) following manufacturer's instructions and vacuum-dried. Dried peptides were dissolved in 50 µl 30% acetonitrile/0.1% TFA and peptide size exclusion (pSEC, Superdex Peptide 3.2/300 column on an ÄKTAmicro system, GE Healthcare) was performed to enrich for crosslinked peptides at a flow rate of 50 µl min⁻¹. Fractions of 50 µl were collected. Fractions containing the crosslinked peptides (1–1.7 ml) were vacuum-dried and dissolved in 2% acetonitrile/0.05% TFA (v/v) for analysis by liquid chromatography with tandem mass spectrometry (LC–MS/MS).

Crosslinked peptides derived from pSEC were analysed as technical duplicates on an Orbitrap Fusion and Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Scientific), respectively, coupled to a Dionex UltiMate 3000 UHPLC system (Thermo Scientific) equipped with an in-house-packed C₁₈ column (ReproSil–Pur 120 C18–AQ, 1.9 µm pore size, 75 µm inner diameter, 30 cm length, Dr. Maisch GmbH). Samples were separated applying the following 58 min gradient: mobile phase A consisted of 0.1% formic acid (v/v), mobile phase B of 80% acetonitrile/0.08% formic acid (v/v). The gradient started at 5% B, increasing to 8% B on Fusion and 15% on Fusion Lumos, respectively, within 3 min, followed by 8–42% B and 15–46% B within 43 min accordingly, then keeping B constant at 90% for 6 min. After each gradient the column was again equilibrated to 5% B for 6 min. The flow rate was set to 300 nl min⁻¹. MS1 spectra were acquired with a resolution of 120,000 in the Orbitrap covering a mass range of 380–1580 *m/z*. Injection time was set to 60 ms and automatic gain control target to 5×10^5 . Dynamic exclusion covered 10 s. Only precursors with a charge state of 3–8 were included. MS2 spectra were recorded with a resolution of 30,000 in the Orbitrap, injection time was set to 128 ms, automatic gain control target to 5×10^4 and the isolation window to 1.6 *m/z*. Fragmentation was enforced by higher-energy collisional dissociation at 30%.

Raw files were converted to mgf format using ProteomeDiscoverer 1.4 (Thermo Scientific, signal-to-noise ratio 1.5, 1,000–10,000 Da precursor mass). For identification of crosslinked peptides, files were analysed by pLink (v. 1.23), pFind group⁸⁷ using BS3 as crosslinker and trypsin as digestion enzyme with maximal two missed cleavage sites. Carbamidomethylation of cysteines was set as a fixed modification, oxidation of methionines as a variable modification. Searches were conducted in combinatorial mode with a precursor mass tolerance of 5 Da and a fragment ion mass tolerance of 20 p.p.m. The used database contained all proteins within the complex. The false discovery rate was set to 0.01. Results were filtered by applying a precursor mass accuracy of ± 10 p.p.m. Spectra of both technical duplicates were combined and evaluated manually. Crosslinking figures were made with XiNet⁸⁸ and the Xlink Analyzer plugin in Chimera⁸⁹. Distances between structured regions were calculated with Xlink Analyzer version 1.1.

Kinase assays. A modified ATP/NADH coupled ATPase assay was used to measure relative rates of ATP hydrolysis of wild-type P-TEFb and D149N P-TEFb⁹⁰. P-TEFb was titrated from 0–1 μM in a final solution containing 3 mM MgCl_2 , 0.1 mM NADH, 0.4% (w/v) pyruvate kinase/lactate dehydrogenase, 1 mM phosphoenolpyruvate, 4 μM GST-RPB1 1593–1970, 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol and 1 mM DTT. Samples (50 μl , final volume) were incubated for 2 min at 30 °C before adding ATP (1 mM final concentration, pH 7.0). The decrease in absorbance at 340 nm, corresponding to NADH oxidation, was measured in 384-well plates (Greiner Bio-One 384 well, clear flat bottom, 781101) over 60 min at 30 °C in a Tecan Infinite Pro M1000 plate reader. The rate of change in absorbance at 340 nm over time was determined from the linear region of the resulting absorbance curves. The experiment was performed three times and error bars represent the standard deviation between the three measurements.

Immunoblotting experiments were performed with GST-RPB1 1593–1970 and *S. scrofa* Pol II treated with wild-type P-TEFb or P-TEFb(D149N). GST-RPB1 1593–1970 (4 μM) was incubated with 100 nM wild-type P-TEFb or P-TEFb(D149N) in a final buffer containing 3 mM MgCl_2 , 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 1 mM ATP pH 7.5 and 1 mM DTT. The reactions were incubated at 30 °C and aliquots were taken at 0, 1, 5, 10, 15 and 30 min after ATP addition. Reactions were quenched by mixing 3 μl of sample with 12 μl of 4x LDS loading buffer (Invitrogen). Samples (2 μl) were run on 4–12% Bis-Tris gels in MES buffer (ThermoFisher Scientific). Proteins were transferred to nitrocellulose membranes (GE Healthcare Life Sciences) and were blocked with 5% (w/v) milk powder in PBS and 0.1% Tween 20 for 1–3 h at room temperature. The membranes were incubated with antibodies against Ser2 (3E8), Ser5 (3E10) (1:14 dilution, gift from D. Eick), and the CTD (MAB10601, MBL International Corporation, 1:1,000) overnight at 4 °C. Antibodies were diluted in 2.5% (w/v) milk in PBS with 0.1% Tween 20. Membranes were washed three times with PBS with 0.1% Tween 20. HRP-conjugated anti-rat secondary antibody (1:5,000) (Sigma-Aldrich A9037) was incubated with the membranes treated with Ser2 and Ser5 antibodies whereas the CTD membrane was treated with HRP conjugated anti-mouse antibody (1:3,000) (Abcam, ab5870) in PBS with 0.1% Tween 20 for 1 h at room temperature. SuperSignal West Pico Chemiluminescent Substrate (ThermoFisher) was used for detection.

S. scrofa Pol II (2.4 μM) was treated with 0.4 μM wild-type P-TEFb for 1 h at 30 °C in a buffer containing 6 mM MgCl_2 , 3 mM ATP pH 7.0, and 1 Roche PhosStop tablet (Sigma-Aldrich) before size-exclusion chromatography. An aliquot of the peak fraction was used for immunoblotting. 0.5 μg of the protein was loaded on 4–12% Bis-Tris SDS-PAGE. Western blotting procedures are identical to those described for the GST-RPB1 CTD. Blots were performed using additional antibodies raised against phosphorylated Ser7 (4E12) and Tyr1 (3D12) (gift from D. Eick) as described above.

Phosphorylation site mapping. NELF, PAF and SPT6 were treated with lambda protein phosphatase during purification to remove phosphorylations that were added by insect cells during protein expression. P-TEFb was incubated with individual factors or with elongation complexes containing PAF, SPT6 and DSIF before size-exclusion chromatography. Protein or fractions from gel-filtration chromatography were applied to NuPAGE 4–12% Bis-Tris SDS-PAGE gels (ThermoFisher Scientific) and stained with InstantBlue (Expediton). Appropriate bands were selected for MS analysis.

Phosphopeptides derived after in-gel digest of the sample were enriched as described previously⁹¹. Enriched phosphopeptides were analysed on a LC-coupled Q-Exactive HF mass spectrometer (ThermoFisher Scientific) under standard chromatography conditions as described⁹¹. The MS raw files were processed by MaxQuant⁹² (version 1.5.2.8) and MS/MS spectra were searched against Uniprot human database with Andromeda⁹³ search engine. Allowed variable modifications included phosphorylation of serine, threonine and tyrosine, methionine oxidation, and carbamidomethylation of cysteine. Sites reported here were present in at least two biological replicates. 80% of the reported sites are found in Phosida⁹⁴ and PhosphoSitePlus⁹⁵ (with the following exceptions: SPT5 S148, S149, T153; SPT6 S1525; NELF-A S244 and all sites for CDC73 and PAF1). Five sites we detected were previously shown to be P-TEFb phosphorylation sites (NELF-E S181; NELF-A T277, S363; SPT5 666, 806)^{10,12,42}.

In vitro kinase assay and mapping of phosphorylation sites in the CTD linker. We were unable to detect P-TEFb specific phosphorylation sites in the linker from our initial MS experiments because the region is devoid of basic residues that are required for tryptic digestion. The CTD linker has a high frequency of hydrophobic residues, which makes it amenable for chymotrypsin digestion. MBP RPB1 1488–1592 was incubated with wild-type P-TEFb or P-TEFb mutant D149N for 30 min at 30 °C under the following conditions: P-TEFb 1 μM , RPB1 1488–1592 36.5 μM , 3 mM ATP pH 7.0, 6 mM MgCl_2 , 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol and 1 mM DTT. The assay was independently conducted twice.

Immediately after each in vitro phosphorylation reaction, proteins were precipitated using chloroform–methanol extraction as described⁹⁶. Protein precipitates

were resuspended in 50 mM ammonium bicarbonate containing 1% Rapigest surfactant (Waters), reduced with DTT and alkylated with iodoacetamide. Residual iodoacetamide was quenched with DTT. The Rapigest concentration was adjusted to 0.1% with 50 mM ammonium bicarbonate and CaCl_2 was added to a final concentration of 2 mM. Proteins were digested at a weight ratio of 75:1 with chymotrypsin (Roche) for 12 h at 25 °C⁹⁷. The digest was then acidified and insoluble material was removed by centrifugation. The peptide mixture was desalted using self-made StageTips⁹⁸ containing Empore C18 solid phase extraction material (3M). Each sample was analysed in duplicate using an Orbitrap Fusion Lumos Tribrid mass spectrometer (ThermoFisher Scientific) coupled to a Dionex UltiMate 3000 nano liquid-chromatography system (ThermoFisher Scientific). Peptides were initially loaded onto a C18-trap column (0.3 \times 5 mm, Dionex) in loading buffer (2% acetonitrile/0.05% TFA) and then separated on an analytical column (self-packed with 1.9 μm ReproSil-Pur C18-AQ material, 30 cm \times 75 μm , Dr. Maisch) at flow rate of 300 nl min⁻¹ using a 90 min multi-step gradient (2% acetonitrile/0.1% formic acid to 48% acetonitrile/0.1% formic acid).

The mass spectrometer was operated in a data-dependent mode to select from a MS survey scan (range: 350–1,550 m/z) the up to 20 most intense peptide precursors with charge states 2–7 for higher-energy collisional dissociation. Spectra were acquired in the Orbitrap at a resolution of 120,000 (MS1) and 30,000 (MS2) with automatic gain control target values of 6×10^5 (MS1) and 1.5×10^4 (MS2) respectively. A dynamic precursor exclusion of 10 s was used. MaxQuant⁹² (version 1.5.2.8) equipped with the Andromeda search engine⁹³ was used to analyse the raw files against a database containing the recombinant protein sequences. Chymotrypsin was selected as protease with cleavage specificity for Trp, Tyr, Phe, Leu and Met. A maximum of two missed cleavage sites was allowed. Precursor and fragment ion tolerances during database search were 4.5 p.p.m. (after internal recalibration) and 20 p.p.m., respectively. Cysteine carbamidomethylation was set as static modification; serine, threonine and tyrosine phosphorylation, methionine oxidation and N-terminal protein acetylation were variable modifications. Label-free quantification was enabled. False discovery rates for peptide-spectrum matches and protein identifications were set to 1%. Phosphorylation sites were filtered for high confidence ($P > 0.75$), further examined manually and only considered relevant, when phosphorylated precursors were identified in both injection and assay replicates. Selected annotated MS2 spectra were exported as vector graphic using the MaxQuant Viewer and for better legibility labels were further modified in Adobe Illustrator CS6 (version 16.0.0).

Full sequence coverage for the RPB1-linker region and near-complete coverage (>88%) for MBP could be obtained for all replicates. Only non-phosphorylated MBP peptides were detected, confirming the specificity of the phosphorylation reaction. Phosphorylated CTD linker peptides were observed after incubation with wild-type P-TEFb (but not with D149N P-TEFb) and phosphorylation sites could be assigned with high confidence (post translational modification (PTM) score >0.98) to six different residues within the CTD linker (S1514, T1518, T1525, T1540, S1584 and S1590). We did not obtain direct evidence for phosphorylation of S1547, which was previously described to co-mediate SPT6 recruitment in yeast⁴⁴, although we noted a considerable intensity decrease for the unphosphorylated counterpart peptide upon incubation with wild-type P-TEFb.

Fluorescence anisotropy binding assays with CTD linker peptide. A 5,6-FAM-labelled peptide corresponding to human RPB1 residues 1521–1552 was purchased from Caslo Aps. The peptide was dissolved in 22% dimethylformamide to a concentration of 1 mM. The peptide (30 μM) was incubated with 1 μM wild-type P-TEFb or P-TEFb(D149N) in buffer containing 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 6 mM MgCl_2 , 3 mM ATP pH 7.0 and 1 mM DTT for 30 min at 30 °C. The reaction was quenched by addition of 10 mM EDTA pH 8.0.

The tSH2 was diluted in half-log steps in protein dilution buffer and mixed with diluted peptide on ice for 5 min. The reaction was diluted to achieve a final volume of 20 μl , final conditions: 10 nM peptide, 100 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol, 1 mM DTT, 1 mM EDTA pH 8.0. The reactions were incubated at room temperature for 20 min in the dark. 17 μl aliquots were removed and transferred to a Greiner 384-well black flat bottom small volume plate. Fluorescence anisotropy was measured at room temperature with an Infinite M1000 Pro plate reader (Tecan) with an excitation wavelength of 470 ± 5 nm, an emission wavelength of 518 ± 5 nm, and a gain of 75. The experiment was performed in triplicate and analysed with GraphPad Prism Version 6. Binding curves were fit using a single site quadratic binding equation as described⁵².

Fluorescence anisotropy binding assays with SPT6 and nucleic acids. SPT6 was dialysed overnight against SPT6 FA buffer (150 mM NaCl, 20 mM Na-HEPES pH 7.4, 10% (v/v) glycerol and 1 mM DTT) in a 20 kDa MWCO Slide-A-Lyzer MINI Dialysis Unit. The protein was then used directly for anisotropy measurements. 25-mer 5′-/6-FAM labelled DNA and RNA oligonucleotides bearing the sequence AAG GGG AGC GGG GGA GGA TAA TAG G (T substituted with U in RNA sequence) were obtained from Integrated DNA Technologies and dissolved in water. SPT6 was serially diluted in half-log steps in SPT6 FA buffer. Nucleic acids

(2.2 μ M, 10 nM final concentration) and SPT6 (4.4 μ M, 0–9.5 μ M final concentration) were mixed on ice and incubated for 5 min. The assay was brought up to a final volume of 22 μ L and incubated at room temperature in the dark for 20 min (final conditions: 30 mM NaCl, 3 mM MgCl₂, 20 mM Na-HEPES pH 7.4, 50 μ M BSA, 5 μ M yeast tRNA and 1 mM DTT). 18 μ L of each solution was transferred to a Greiner 384 Flat Bottom Black Small volume plate. Fluorescence anisotropy was measured and analysed as above but with a gain of 70 and an emission wavelength of 518 \pm 20 nm.

Pull-down experiments. MBP-RPB1 1488–1970 and MBP RPB1 1593–1970 (5 μ M) were incubated with 0.4 μ M wild-type or P-TEFb(D149N) mutant in pulldown buffer (100 mM NaCl, 20 mM Na-HEPES pH 7.4, 4% (v/v) glycerol, 1 mM DTT and 3 mM MgCl₂) for 30 min at 30 °C. The CTD constructs were incubated with amylose beads for 10 min further. The beads were washed three times with pulldown buffer to remove P-TEFb and ATP. Full-length SPT6 or SPT6 Δ tSH2 were then added at a final concentration of 7.5 μ M. The reactions were incubated at 30 °C for 15 min and washed three times with pulldown buffer. The MBP tag was eluted from the beads by applying pulldown buffer with 116 mM maltose to the beads. 20 μ L of the eluted sample was applied to a 4–12% SDS–PAGE and stained with Coomassie blue.

Crystal structure determination of SPT6 tSH2 domain. Frozen tSH2 protein was thawed and applied to a Superdex S200 increase 10/300 column equilibrated in SE buffer. Peak fractions were pooled, concentrated, and dialysed into a buffer containing 100 mM NaCl, 20 mM HEPES pH 7.4, and 1 mM TCEP pH 7.0 for 16 h at 4 °C. Initial crystals of tSH2 were obtained by hanging-drop vapour diffusion crystallization at 293 K by mixing 1 μ L of protein solution and 1 μ L of reservoir solution containing 100 mM Bis-Tris pH 5.5, 200 mM MgCl₂ and 21–25% (v/v) PEG 3350. Larger crystals were obtained by micro-seeding. A drop with initial crystal hits was transferred to a microcentrifuge tube containing one glass bead (Bead for seeds, Jena Biosciences) and 50 μ L of solution (100 mM Bis-Tris pH 5.5, 200 mM MgCl₂ and 25% (v/v) PEG 3350). The seed stock was vortexed extensively and diluted 1:1,000. Crystals used for data collection were grown using the hanging-drop vapour diffusion technique with a 0.5:0.5:1 ratio of protein solution, seed solution and reservoir solution (100 mM Bis-Tris pH 5.5, 200 mM MgCl₂ and 25% (v/v) PEG 3350) at 293 K. For collection, crystals were exchanged into cryo-protectant solution (100 mM Bis-Tris pH 5.5, 200 mM MgCl₂, 25% (v/v) glycerol and 25% (v/v) PEG 3350), and flash-frozen in liquid nitrogen.

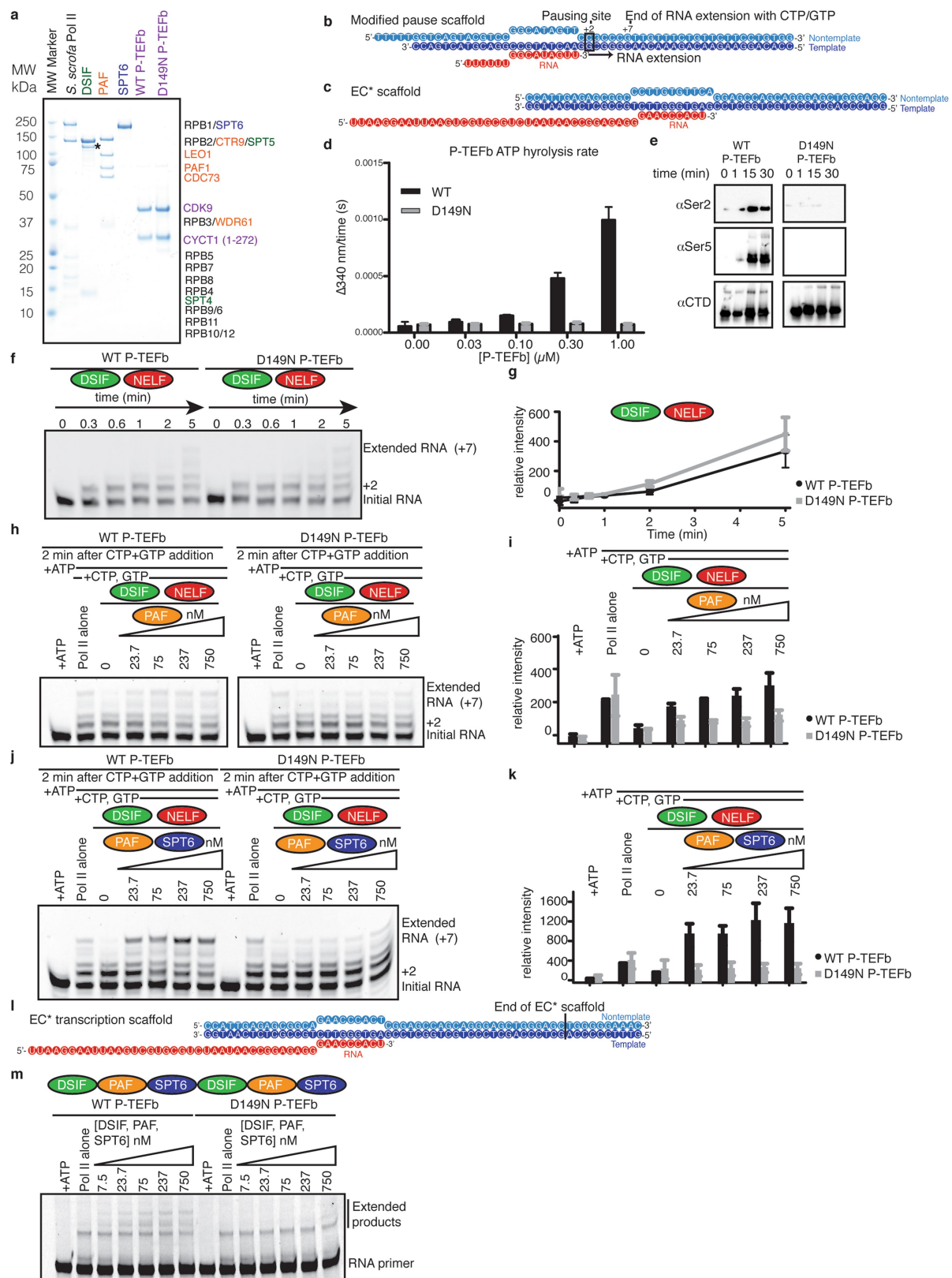
Diffraction data were collected at beamline PXII of the Swiss Light Source at the Paul Scherrer Institute⁹⁹. The native dataset was collected at a wavelength of 0.999 Å. Diffraction images were processed with XDS¹⁰⁰. The structure was solved with molecular replacement in Phaser¹⁰¹ using a polyaniline model of the *S. cerevisiae* Spt6 tSH2 domain (PDB ID: 3PSJ). Refinement was performed using Phenix.Refine¹⁰² applying riding hydrogens. The human tSH2 structure is nearly identical to the yeast structure except for an N-terminal loop that is involved in crystal packing and adopts an alternative conformation (Extended Data Fig. 8b). The final model was refined to an $R_{\text{work}}/R_{\text{free}}$ of 19.0%/22.3%. MolProbity⁸⁴ analysis showed that 98.61% of the residues reside in the most-favoured regions of the Ramachandran plot, and 1.39% fell in allowed regions. None of the residues fell in disallowed regions.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The electron density reconstructions and the final EC* model were deposited with the Electron Microscopy Data Base (EMDB) under accession codes EMD-0030, EMD-0031, EMD-0032, EMD-0033, EMD-0034, EMD-0035, EMD-0036 and EMD-0037, and with the Protein Data Bank (PDB) under accession code 6GMH. The tSH2 domain model was deposited with the PDB under accession code 6GME. Source data for Figs. 1a and 5a, Extended Data Figs. 1a, e–k, m, 2a–j, l, m, 9a, and 10g–k are found in Supplementary Figs 1, 2 and Supplementary Table 8.

51. Gradia, S. D. et al. MacroBac: new technologies for robust and efficient large-scale production of recombinant multiprotein complexes. *Methods Enzymol.* **592**, 1–26 (2017).
52. Vos, S. M. et al. Architecture and RNA binding of the human negative elongation factor. *eLife* **5**, e14981 (2016).
53. Kapust, R. B. & Waugh, D. S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674 (1999).
54. Sydow, J. F. et al. Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell* **34**, 710–721 (2009).
55. Sidorenkov, I., Komissarova, N. & Kashlev, M. Crucial role of the RNA:DNA hybrid in the processivity of transcription. *Mol. Cell* **2**, 55–64 (1998).
56. Larson, M. H. et al. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042–1047 (2014).
57. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
58. Komissarova, N., Kireeva, M. L., Becker, J., Sidorenkov, I. & Kashlev, M. Engineering of elongation complexes of bacterial and yeast RNA polymerases. *Methods Enzymol.* **371**, 233–251 (2003).
59. Xu, J. et al. Structural basis for the initiation of eukaryotic transcription-coupled DNA repair. *Nature* **551**, 653–657 (2017).
60. Kireeva, M. L., Komissarova, N. & Kashlev, M. Overextended RNA:DNA hybrid as a negative regulator of RNA polymerase II processivity. *J. Mol. Biol.* **299**, 325–335 (2000).
61. Kireeva, M. L., Komissarova, N., Waugh, D. S. & Kashlev, M. The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J. Biol. Chem.* **275**, 6530–6536 (2000).
62. Zheng, S. Q. et al. MotionCorr2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
63. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
64. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
65. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
66. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
67. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
68. Plaschka, C. et al. Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* **518**, 376–380 (2015).
69. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
70. Meka, H., Werner, F., Cordell, S. C., Onesti, S. & Brick, P. Crystal structure and RNA binding of the Rpb4/Rpb7 subunits of human RNA polymerase II. *Nucleic Acids Res.* **33**, 6435–6444 (2005).
71. Afonine, P. V., Headd, J. J., Terwilliger, T. & Adams, P. D. New tool: phenix.refine. *Comput. Crystallogr. Newsl.* **4**, 43–44 (2013).
72. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526–W531 (2004).
73. Halbach, F., Reichelt, P., Rode, M. & Conti, E. The yeast ski complex: crystal structure and RNA channeling to the exosome complex. *Cell* **154**, 814–826 (2013).
74. Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59**, 467–475 (2005).
75. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
76. Karpenhalli, M. R., Lupas, A. N. & Söding, J. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics* **8**, 2 (2007).
77. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
78. Xu, Y. et al. Architecture of the RNA polymerase II–Paf1C–TFIIS transcription elongation complex. *Nat. Commun.* **8**, 15741 (2017).
79. Rozenblatt-Rosen, O. et al. The parafibromin tumor suppressor protein is part of a human Paf1 complex. *Mol. Cell. Biol.* **25**, 612–620 (2005).
80. Yart, A. et al. The HRPT2 tumor suppressor gene product parafibromin associates with human PAF1 and RNA polymerase II. *Mol. Cell. Biol.* **25**, 5052–5060 (2005).
81. Amrich, C. G. et al. Cdc73 subunit of Paf1 complex contains C-terminal Ras-like domain that promotes association of Paf1 complex with chromatin. *J. Biol. Chem.* **287**, 10863–10875 (2012).
82. Cao, Q.-F. et al. Characterization of the human transcription elongation factor Rtf1: evidence for nonoverlapping functions of Rtf1 and the Paf1 complex. *Mol. Cell. Biol.* **35**, 3459–3470 (2015).
83. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
84. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
85. Dolinsky, T. J. et al. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522–W525 (2007).
86. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**, 10037–10041 (2001).
87. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
88. Combe, C. W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol. Cell. Proteomics* **14**, 1137–1147 (2015).
89. Kosinski, J. et al. Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J. Struct. Biol.* **189**, 177–183 (2015).
90. Kiianitsa, K., Solinger, J. A. & Heyer, W.-D. NADH-coupled microplate photometric assay for kinetic studies of ATP-hydrolyzing enzymes with low and high specific activities. *Anal. Biochem.* **321**, 266–271 (2003).

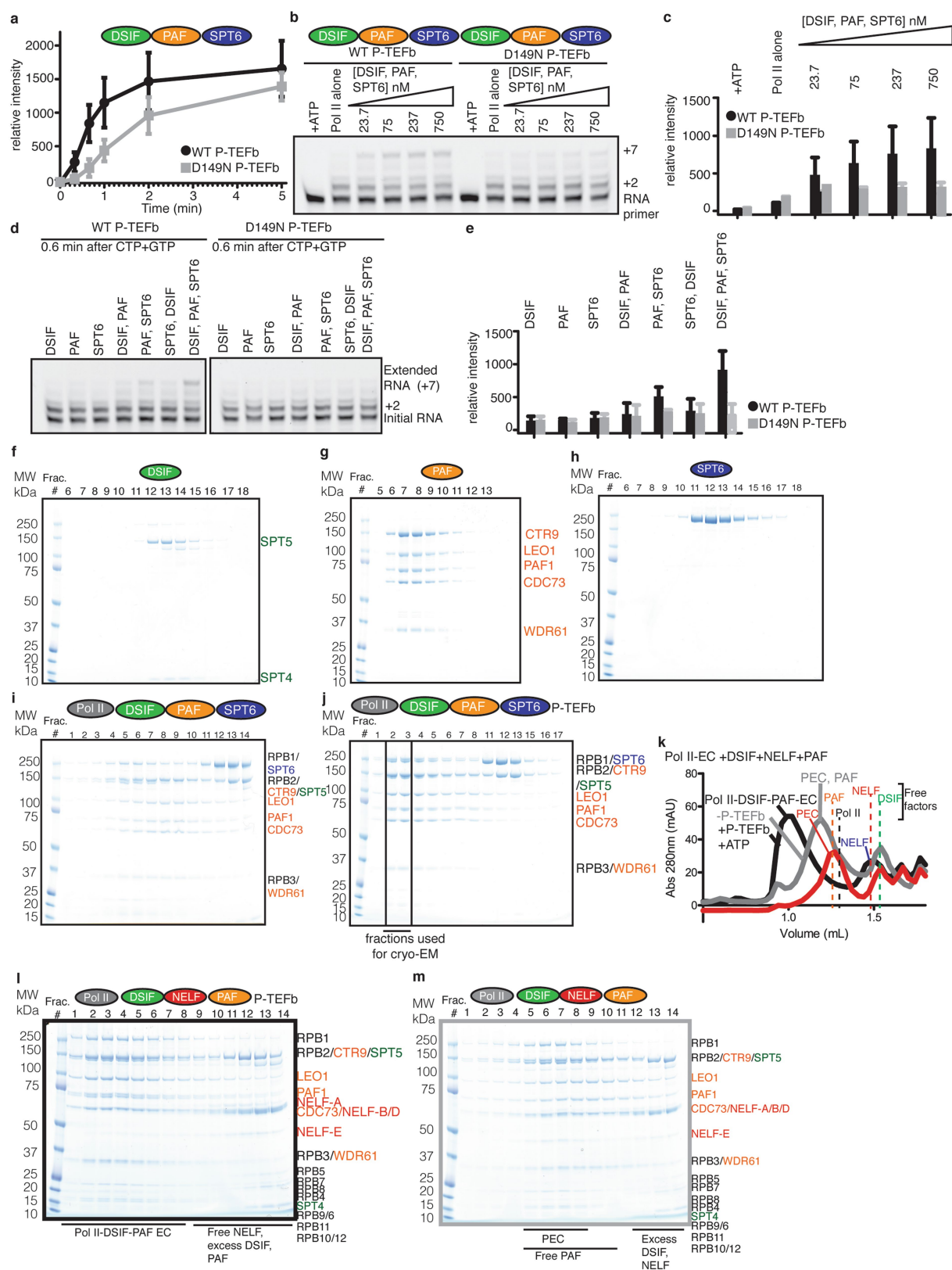
91. Oellerich, T. et al. SLP-65 phosphorylation dynamics reveals a functional basis for signal integration by receptor-proximal adaptor proteins. *Mol. Cell. Proteomics* **8**, 1738–1750 (2009).
92. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
93. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
94. Gnad, F. et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8**, R250 (2007).
95. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
96. Wessel, D. & Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–143 (1984).
97. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11**, 993–1006 (2016).
98. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
99. Fuchs, M. R. et al. D3, the new diffractometer for the macromolecular crystallography beamlines of the Swiss Light Source. *J. Synchrotron Radiat.* **21**, 340–351 (2014).
100. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
101. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
102. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
103. Sun, M., Larivière, L., Deng, S., Mayer, A. & Cramer, P. A tandem SH2 domain in transcription elongation factor Spt6 binds the phosphorylated RNA polymerase II C-terminal repeat domain (CTD). *J. Biol. Chem.* **285**, 41597–41603 (2010).
104. Diebold, M.-L. et al. Noncanonical tandem SH2 enables interaction of elongation factor Spt6 with RNA polymerase II. *J. Biol. Chem.* **285**, 38389–38398 (2010).
105. Schilbach, S. et al. Structures of transcription pre-initiation complex with TFIID and Mediator. *Nature* **551**, 204–209 (2017).
106. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
107. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Protein preparation and phosphorylation activity of P-TEFb and RNA extension assays. **a**, Quality of purified proteins used in this study (0.9 μ g protein per lane). All proteins were purified at least twice. The representative gel was run twice. The asterisk denotes a SPT5 N-terminal degradation product. **b**, Nucleic acid scaffold used for RNA extension assays, termed the modified pause scaffold. **c**, Nucleic acid scaffold used for analytical gel filtration and for cryo-EM analysis, termed the EC* scaffold. **d**, P-TEFb kinase activity using a coupled ATP/NADH assay. Bars correspond to the absolute change in absorbance at 340 nm as a function of time. Error bars represent the standard deviation between three individual experiments. Each bar corresponds to the mean of three individual experiments. **e**, P-TEFb (100 nM) was incubated with GST-RPB1 CTD for different amounts of time as indicated. Membranes were incubated with antibodies that recognize phospho-Ser2 (3E10), phospho-Ser5 (3E8), or the CTD (MAB10601). Similar experiments were performed at least three times for the wild-type enzyme. The western blot for the D149N mutant was performed once. **f**, Pol II (75 nM) was incubated with wild-type P-TEFb or P-TEFb(D149N) (100 nM) and DSIF and NELF (150 nM). Reactions were quenched at various time points after the addition of GTP and CTP (10 μ M). The experiment was performed three times. **g**, Quantification of extended RNA products in **f**. Points are the mean of three individual experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **h**, Pol II (75 nM) was incubated with the modified pause scaffold (50 nM) (Extended Data Fig. 1b),

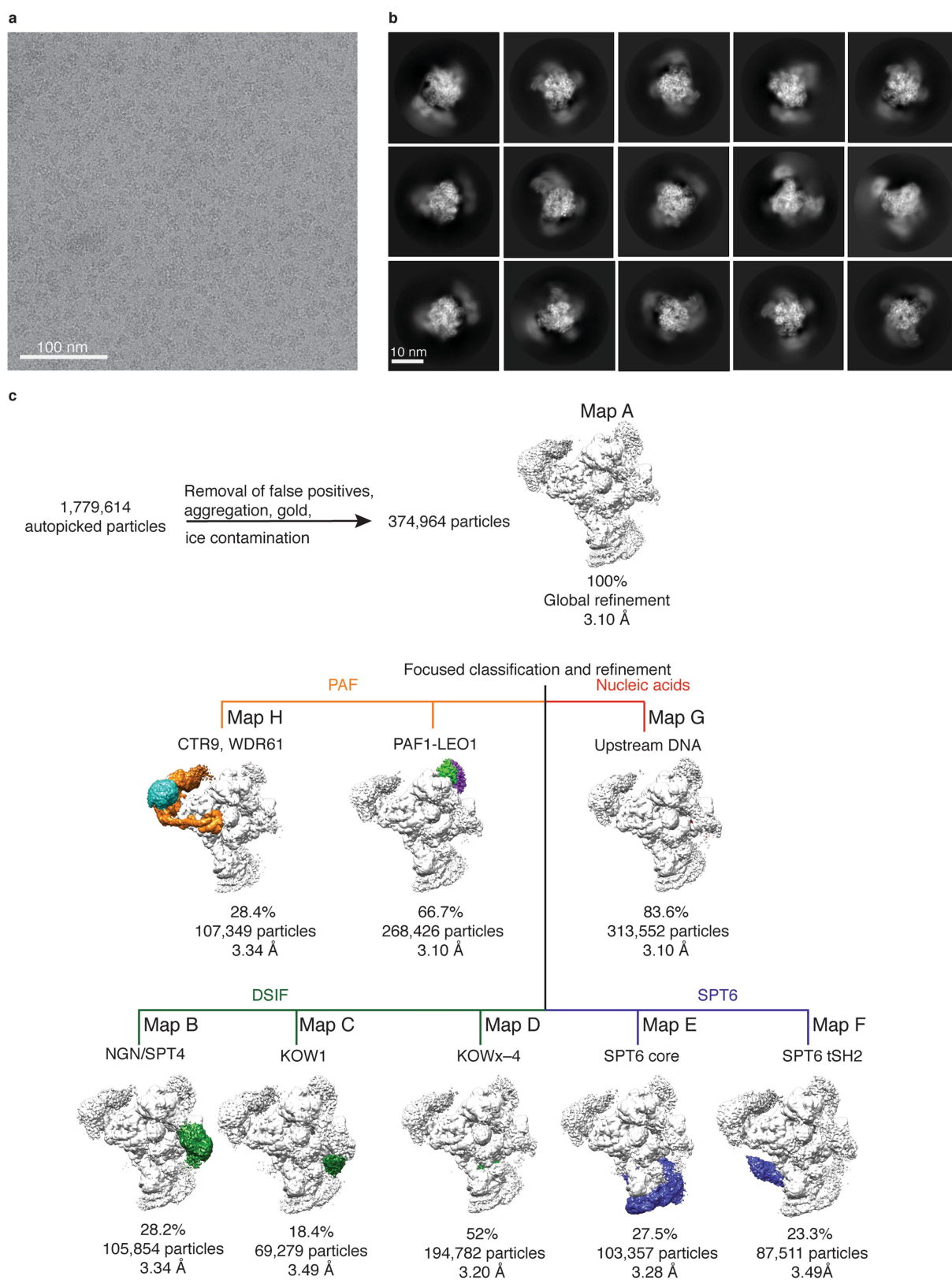
wild-type (WT) P-TEFb or inactive P-TEFb(D149N) (100 nM) and ATP (1 mM) (all lanes), and DSIF and NELF (150 nM). PAF was titrated into the reactions. The reactions were quenched 2 min after the addition of CTP and GTP (10 μ M). Positions for a consensus pausing site (+2) and extended RNA (+7) are marked. RNA extension is incomplete because only a fraction of Pol II molecules assemble on the scaffold. The experiment was performed twice. **i**, Quantification of extended RNA products in **h**. Points are the mean of two individual experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **j**, Pol II (75 nM) was incubated with DSIF and NELF (150 nM) and wild-type P-TEFb or P-TEFb(D149N) (100 nM). PAF and SPT6 were titrated into the reactions. Reactions were quenched 1 min after the addition of GTP and CTP (10 μ M). The experiment was performed three times. **k**, Quantification of extended RNA products in **j**. Points are the mean of three individual experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **l**, Nucleic acid scaffold used for RNA extension assays, termed the EC* transcription scaffold. **m**, RNA extension assays performed on the EC* transcription scaffold (50 nM). Pol II (75 nM) was incubated with elongation factors (7.5–750 nM) (DSIF, PAF, SPT6), active P-TEFb or inactive P-TEFb(D149N) (100 nM) and 1 mM ATP for 15 min. Reactions were quenched 1 min after the addition of GTP, CTP and UTP. Experiments were performed three times. A large fraction of RNA primer remains owing to incomplete assembly of the elongation complex (see Methods for more details).



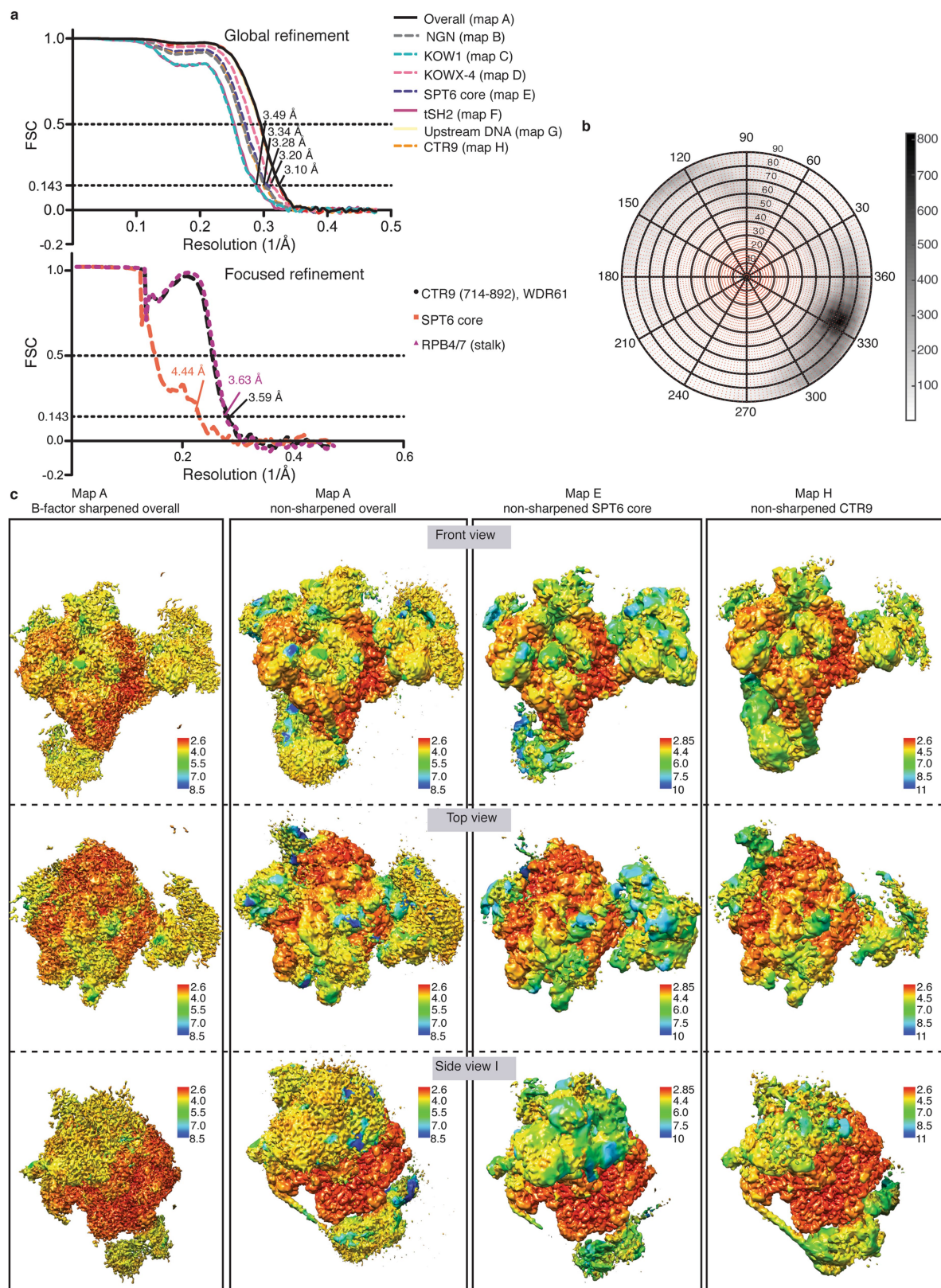
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | P-TEFb activity enables PAF to displace NELF and EC* formation. **a**, Quantification of extended RNA products in Fig. 1a. Points are the mean of three individual experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **b**, PAF, DSIF and SPT6 (23.7–750 nM) were titrated against Pol II (75 nM) and wild-type P-TEFb or P-TEFb(D149N). Reactions were quenched 1 min after the addition of GTP and CTP (10 μ M). The experiment was performed three times. **c**, Quantification of extended RNA products in **b**. Points are the mean of three individual experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **d**, Elongation factors (75 nM) were incubated with P-TEFb (100 nM) and ATP (1 mM). Reactions were quenched after 0.6 min after the addition of GTP and CTP (10 μ M). The experiment was performed three times. **e**, Quantification of extended RNA products in **d**. Points are the mean of three individual

experiments and error bars represent the standard deviation between replicates. Source data: Supplementary Table 8. **f–j**, SDS–PAGE analysis of size-exclusion chromatography fractions. The Pol II elongation complex was formed on the EC* scaffold. All experiments were performed at least twice. **f**, DSIF; **g**, PAF; **h**, SPT6; **i**, Pol II elongation complex, DSIF, PAF, SPT6; **j**, Pol II elongation complex, DSIF, PAF, SPT6, P-TEFb and ATP. Fractions used for cryo-EM are indicated. **k**, NELF is released from Pol II when PAF, wild-type P-TEFb and ATP are present as assessed by size-exclusion chromatography. Curves from the PEC and the PEC plus PAF are shown as a reference. The Pol II elongation complex was formed on the EC* scaffold. Each experiment was performed at least twice. **l**, SDS–PAGE analysis of size-exclusion chromatography fractions from the formation of PEC with PAF, P-TEFb and ATP. The experiment was performed twice. **m**, SDS–PAGE analysis of size-exclusion chromatography fractions from the formation of PEC with PAF. The experiment was performed twice.



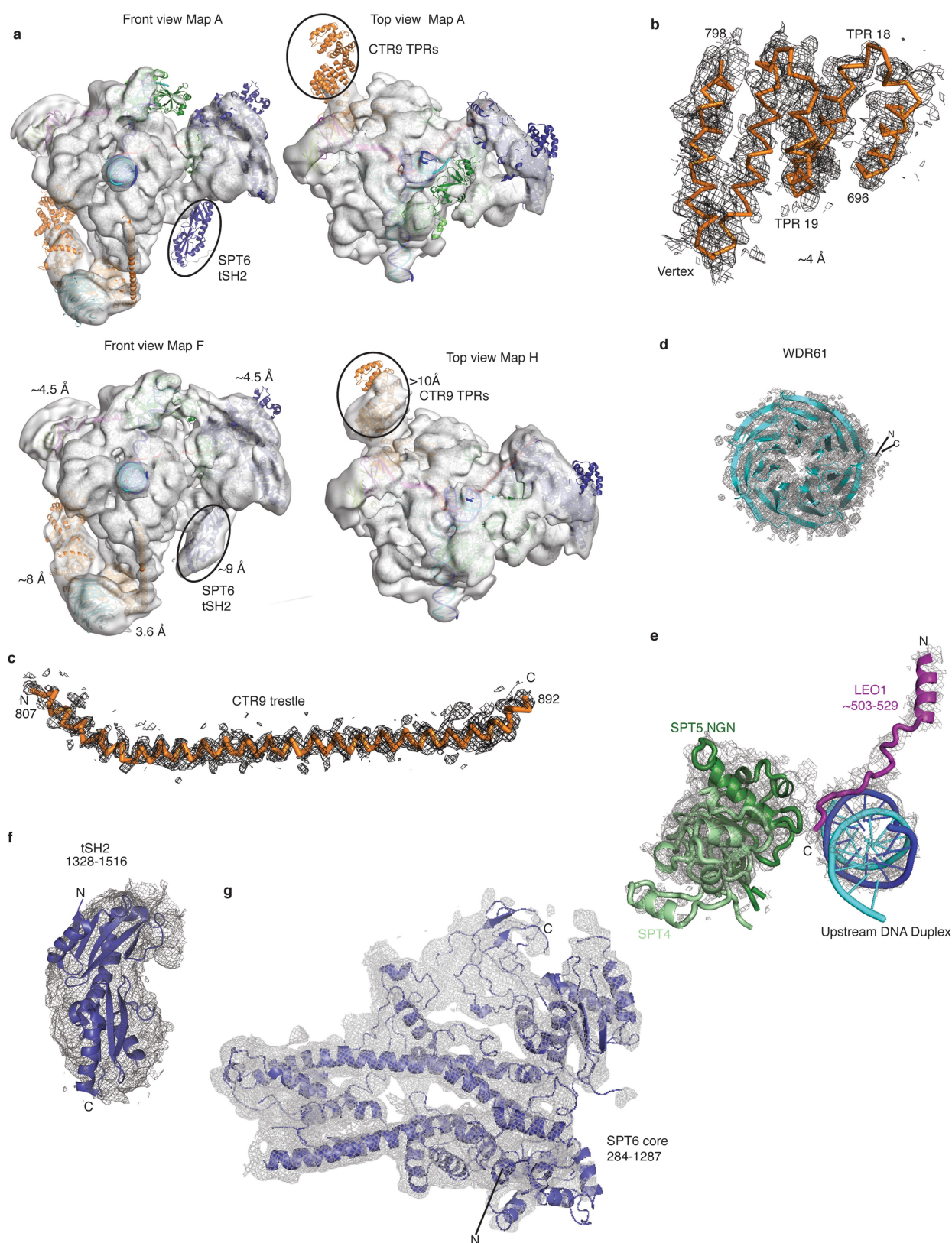
Extended Data Fig. 3 | Cryo-EM data collection and processing. **a**, Representative micrograph of the EC* shown at a defocus of $-2.5\ \mu\text{m}$. Representative of 20,198 micrographs. **b**, Representative 2D classes of EC* particles. **c**, Classification tree for data processing.



Extended Data Fig. 4 | Quality and resolution of cryo-EM data.

a, Estimate of average resolution. Lines indicate the Fourier shell correlation (FSC) between the half maps of the reconstruction. **b**, Angular distribution of particles from overall refinement. Red dots indicate the

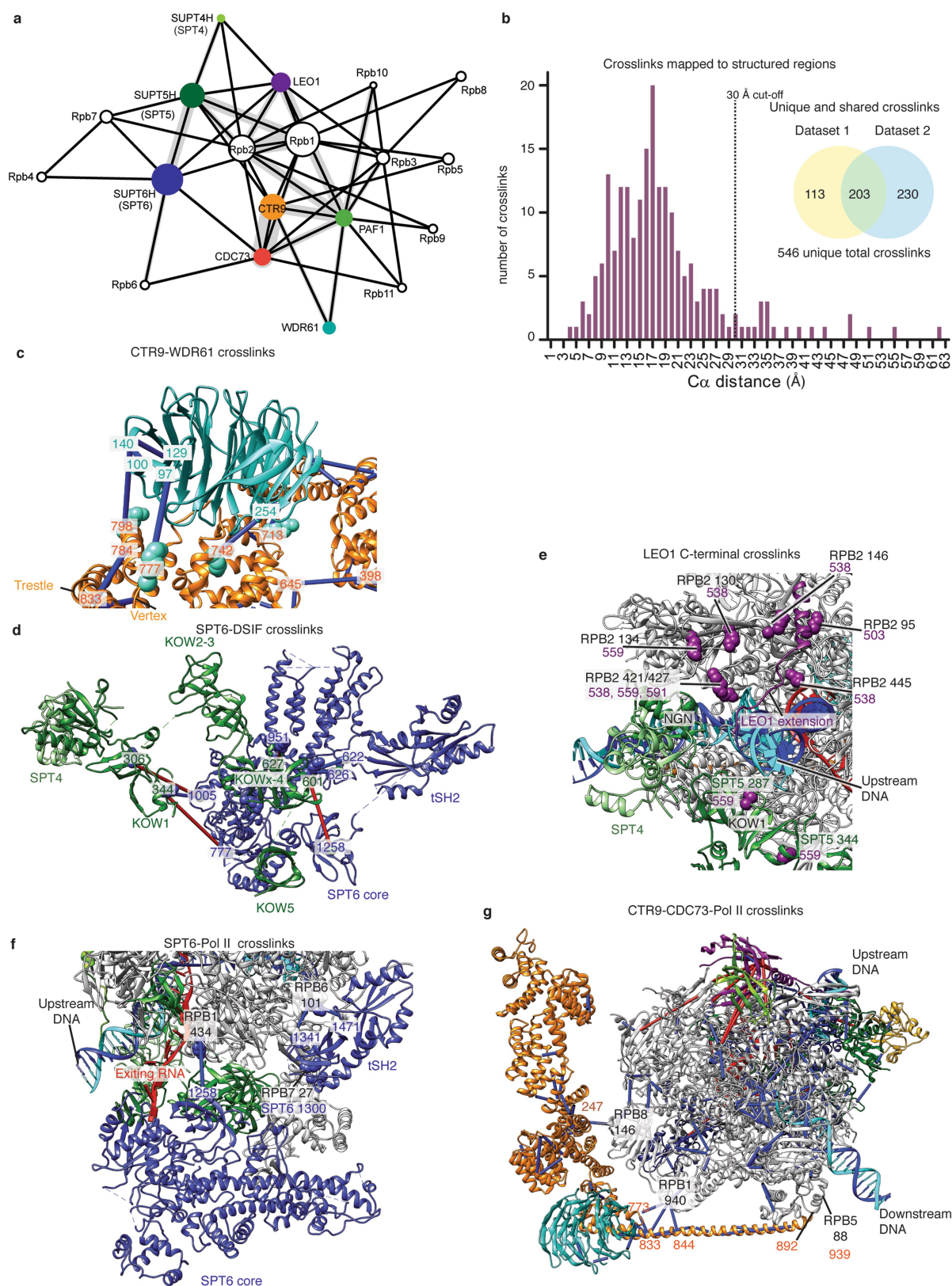
presence of at least one particle image within $\pm 1^\circ$. **c**, Reconstructions of EC* as coloured by local resolution. The overall reconstruction is shown with B-factor-sharpened and non-sharpened maps. The globally refined maps E and H are shown as non-B-factor-sharpened maps.



Extended Data Fig. 5 | Fits of the EC* model in representative densities.

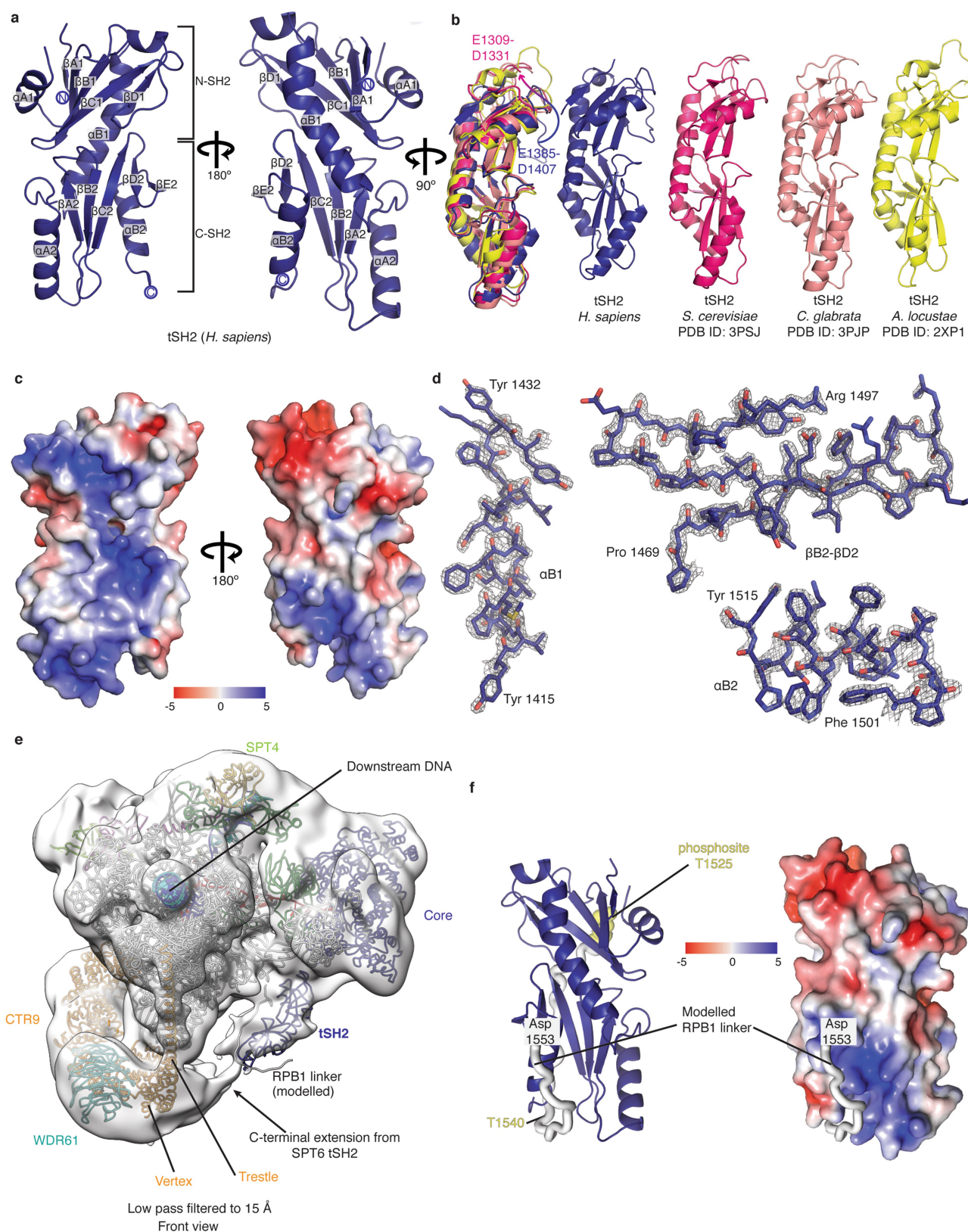
a, EC* fit in electron density (map A) contoured to 12 Å. Black ovals indicate regions where electron density was weak. Map F and map H are shown to indicate the improvement after focused classification and

refinement. **b–f**, Electron density for various elements of the EC* shown as grey mesh. **b**, CTR9 vertex and TPRs 18–19, map H. **c**, CTR9 trestle helix, map H. **d**, WDR61, map H. **e**, C terminus of LEO1 and upstream DNA, map G. **f**, tSH2 crystal structure, map F. **g**, Core of SPT6, map E.



Extended Data Fig. 6 | Crosslinking-mass spectrometry analysis.
a, Overview of crosslinks obtained with BS3 in EC*. Connecting line thickness signifies the number of crosslinks obtained between subunits.
b, Histogram of unique crosslinks and distances between C α pairs that were mapped onto our structure. A dotted black line marks the 30 Å distance cutoff for BS3. The Venn diagram compares unique crosslinks

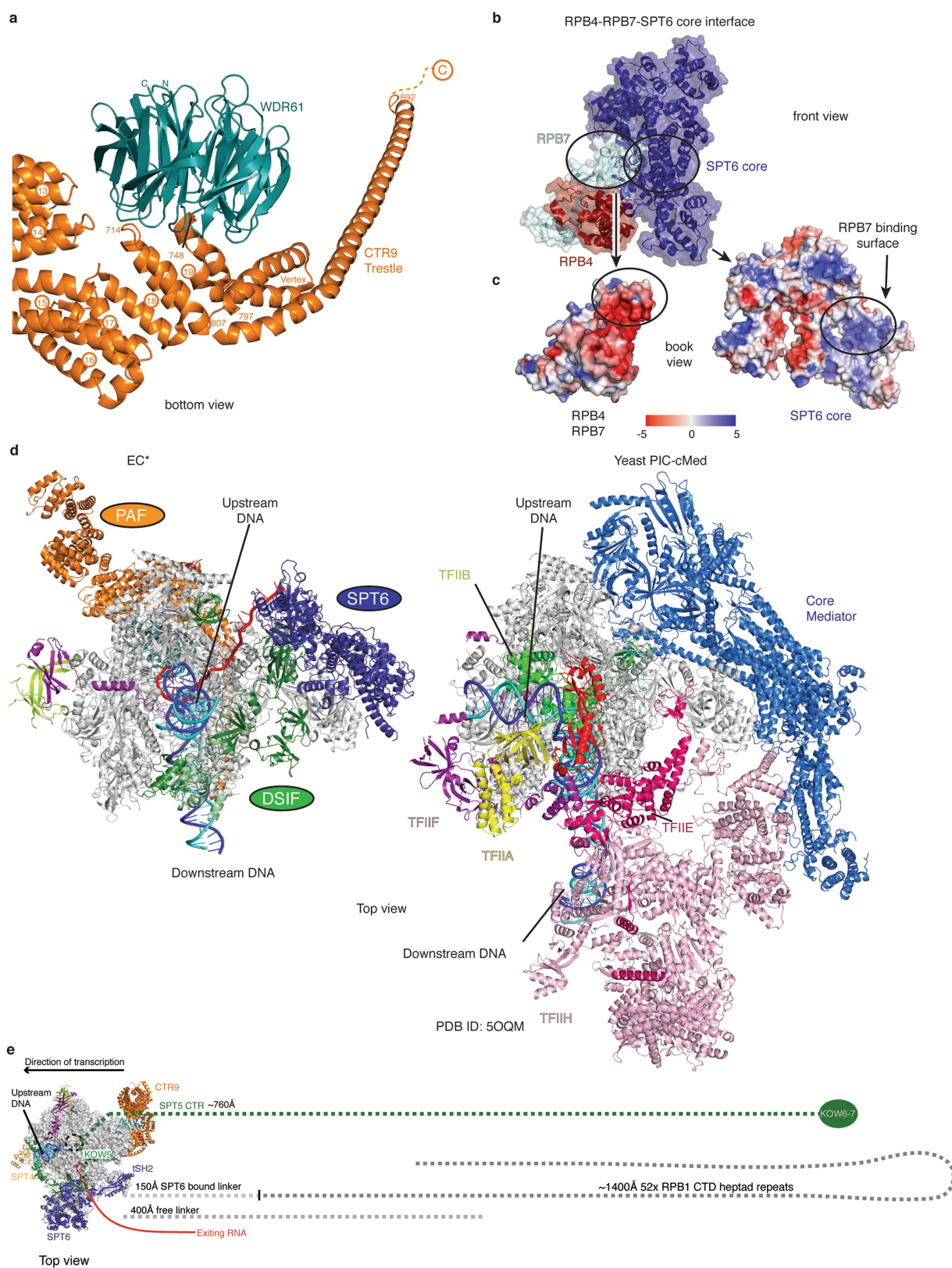
between two biological replicates. **c–g**, Crosslinks mapped onto the final model. Residues involved in crosslinks are shown as spheres. Coloured rods connecting residues signify permitted (blue) or non-permitted (red) crosslinking distances. **c**, WDR61 and CTR9. **d**, DSIF KOW1 and KOWx-4 domains and SPT6. **e**, A C-terminal extension of LEO1, NGN and KOW1 domain of SPT5 and RPB2. **f**, SPT6 and Pol II. **g**, CTR9 and Pol II.



Extended Data Fig. 7 | Crystal structure of SPT6 tSH2 and associated electron microscopy densities. **a**, Cartoon model of human tSH2 crystal structure shown in two different views. **b**, Human SPT6 tSH2 is

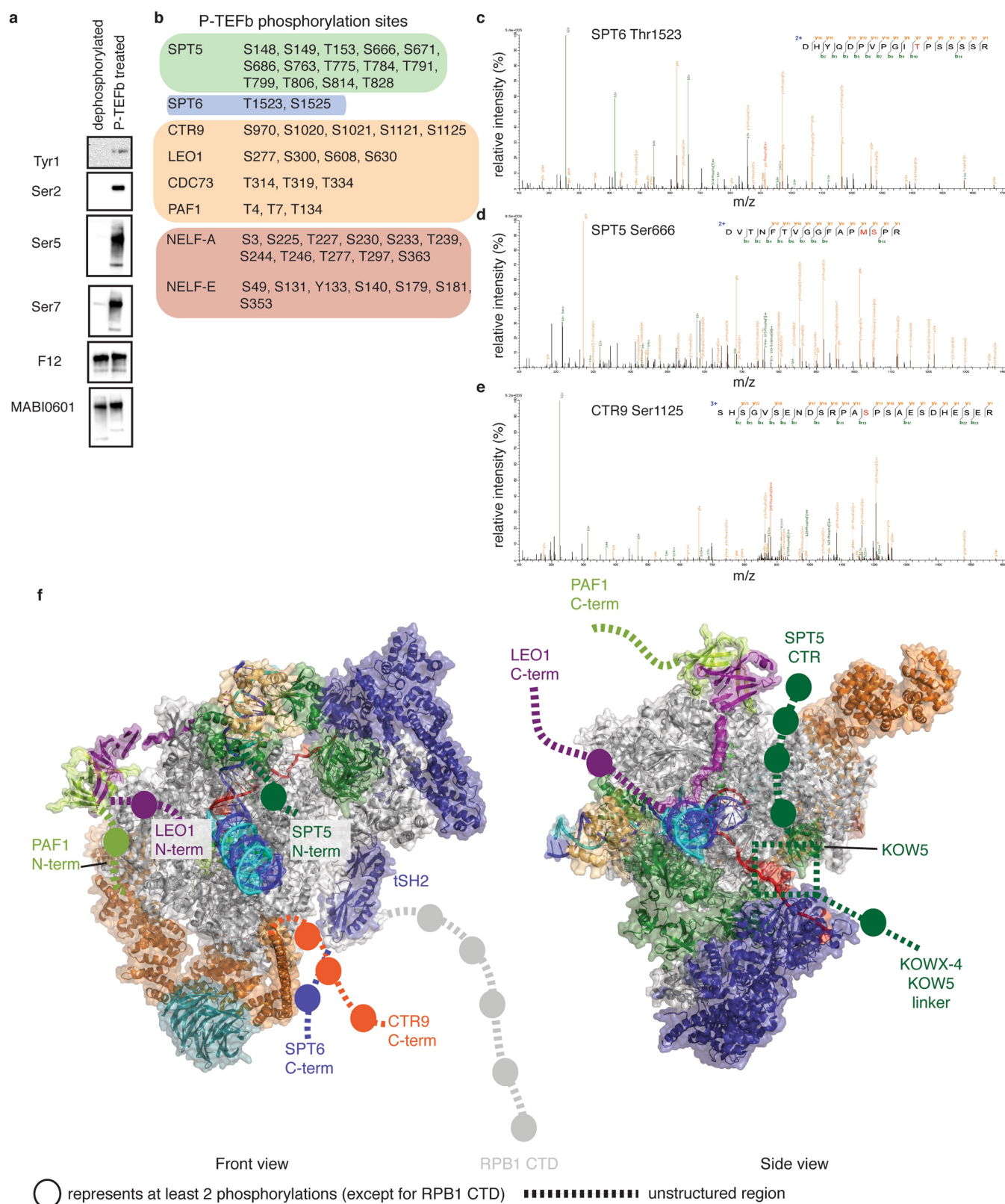
structurally similar to previously obtained SPT6 tSH2 structures from *S. cerevisiae*²⁹ (PDB ID: 3PSJ) (hot pink), *Candida glabrata*¹⁰³ (PDB ID: 3PJP) (grey), and *Antonospora locustae*¹⁰⁴ (PDB ID: 2XP1) (peach). **c**, Surface charge representation of the human SPT6 tSH2.

d, Representative electron density from the crystal structure of tSH2. $2F_o - F_c$ maps contoured at 2σ are shown for several regions of the tSH2 crystal structure. **e**, 15 Å low-pass filtered map E. The C-terminal density of SPT6 extends to CTR9. **f**, Alternative view to that shown in Fig. 5b. Two P-TFEB phosphorylation sites are demarcated (T1525, T1540). The T1540 site was not observed in the yeast linker that was used for crystallization. The CTD linker is modelled.



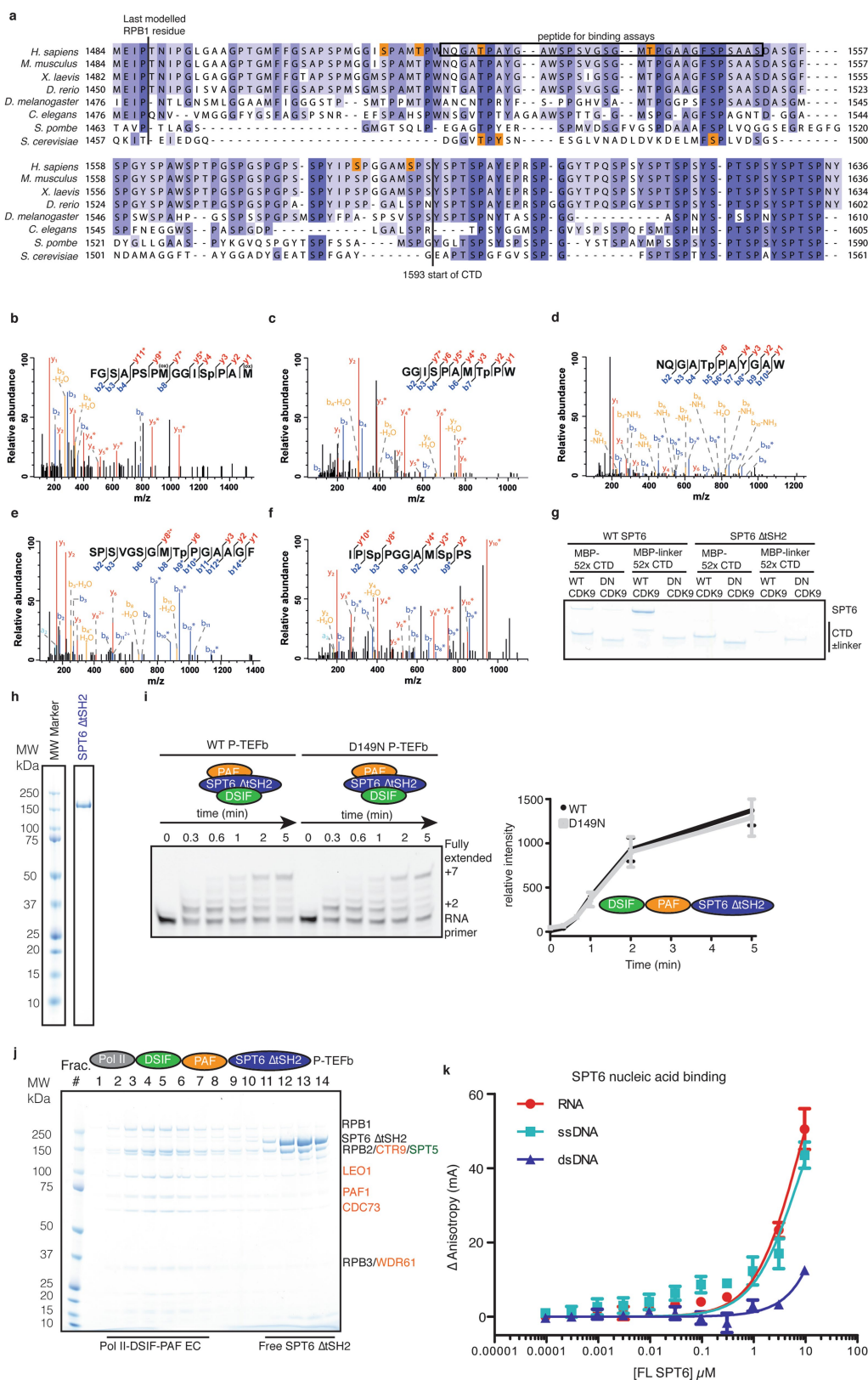
Extended Data Fig. 8 | Features of EC* and comparisons to other structures. **a**, WDR61 is anchored by the vertex and TPRs 13, 18 and 19. **b**, **c**, SPT6 binds to the C1–C3 sheets of RPB7. **b**, Surface representation of the association of SPT6 with the RPB4–RPB7 stalk (RPB4, red; RPB7, cyan). **c**, Book view of **b**. RPB4–RPB7 and SPT6 are coloured according to

surface charge (blue, positive; red, negative). **d**, Comparison of initiation factor and elongation factor binding sites. The yeast preinitiation complex bound to core mediator (PIC-cMed)¹⁰⁵ (PDB ID: 5OQM) was aligned with the EC* Pol II core. **e**, Model of RNA, CTD and CTR paths extending from the EC*.



Extended Data Fig. 9 | EC* is highly phosphorylated. **a**, *S. scrofa* Pol II CTD P-TEFb phosphorylations assessed by western blot using antibodies raised against phospho-Tyr1 (3D12), phospho-Ser2 (3E10), phospho-Ser5 (3E8) and phospho-Ser7 (4E12) or the RPB1 body (F12) or CTD (MABI0601). Experiments with the phospho-antibodies were performed twice. The RPB1 body and CTD antibody experiments were performed

once. **b**, Phosphorylation sites as determined by mass spectrometry. The experiment was performed two or more times with each protein. The reported sites were found in at least two independent replicates. **c–e**, Representative mass spectra. **f**, Phosphorylations map to flexible regions of the EC*. Spheres and dotted lines represent two phosphorylations and flexible regions, respectively.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | P-TEFb phosphorylates the CTD linker and SPT6 tSH2 required for association with EC*. **a**, Sequence alignment of the CTD linker from various species generated in Mafft¹⁰⁶ and visualized in Jalview¹⁰⁷ (*S. cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Xenopus laevis*, *Mus musculus* and *Homo sapiens*). Blue columns represent regions sharing sequence identity. Orange boxes represent phosphorylation sites reported here or those obtained previously in yeast⁴⁴. **b–f**, Representative MS2 spectra of P-TEFb phosphorylated CTD linker peptides. Spectra are representative of two biological replicates. RPB1 residues serine 1514 (**b**; precursor m/z 759.804, $z = +2$, corresponding RPB1 residues 1503–1517), threonine 1518 (**c**; precursor m/z 548.730, $z = +2$, RPB1 residues 1511–1520), threonine 1525 (**d**; precursor m/z 608.245, $z = +2$, RPB1 residues 1521–1531), threonine 1540 (**e**; precursor m/z 701.789, $z = +2$, RPB1 residues 1532–1546) as well as serine 1584 and serine 1590 (**f**; precursor m/z 580.708, $z = +2$, RPB1 residues 1582–1592) are phosphorylated by P-TEFb in vitro. The sequence of the corresponding phosphorylated chymotryptic precursor peptide is shown with all identified b-ions (blue) and y-ions (red). Asterisks indicate neutral loss

of phosphoric acid (H_3PO_4 , $\Delta 97.98$ Da), which is commonly observed for phosphoserine- and phosphothreonine-containing peptides upon HCD fragmentation. Additionally, peaks corresponding to neutral loss of ammonia (NH_3 , $\Delta 17.03$ Da) or water (H_2O , $\Delta 18.01$ Da) are labelled in orange. **g**, Pulldowns performed with full-length SPT6 and SPT6 $\Delta tSH2$ and MBP-RPB1 CTD constructs in the presence of wild-type P-TEFb or P-TEFb(D149N). The gel is representative of two independent experiments. **h**, Quality of purified SPT6 $\Delta tSH2$ (1–1297) (0.9 μg). **i**, Time-course transcription assay with SPT6 $\Delta tSH2$, PAF, DSIF (75 nM) and wild-type P-TEFb or P-TEFb(D149N). The gel is representative of three independent experiments. **j**, Size-exclusion chromatography experiment as performed in Extended Data Fig. 1. SPT6 $\Delta tSH2$ does not stably associate with the EC*. The experiment was performed twice. **k**, Nucleic acid association with full-length SPT6. Binding to single-stranded DNA (cyan), double-stranded DNA (blue) or RNA (red) was assessed by fluorescence anisotropy. Error bars reflect the standard deviation between three experimental replicates. Points represent the mean of three experimental replicates.

Extended Data Table 1 | Components of the EC*

Component	Subunit	Construct residues (aa) / scaffold length (nt)	Mass (kDa)	UniProt/Genbank identifier
	RPB1	1-1970	217.2	XP_020923484.1
	RPB2	1-1174	133.8	XP_003129085.4
	RPB3	1-275	31.4	XP_003355849.1
	RPB4	1-142	16.3	XP_020932152.1
	RPB5	1-210	24.6	XP_003354010.1
<i>S. scrofa</i> Pol II	RPB6	1-127	14.4	XP_003481589.1
	RPB7	1-172	19.2	XP_013849657.1
	RPB8	1-150	17.1	NP_001230270.1
	RPB9	1-125	14.5	NP_001192333.1
	RPB10	1-67	7.6	XP_003122432.1
	RPB11	1-117	13.2	XP_003124442.2
	RPB12	1-58	7.0	XP_003355060.1
	CDC73	1-531	60.6	Q6P1J9-1
	CTR9 ^a	1-1173	133.5	Q6PD62-1
PAF1c	LEO1	1-666	75.4	Q8WVC0-1
	PAF1	1-531	60.0	Q8N7H5-1
	WDR61	1-305	33.6	Q9GZS3-1
	SPT4 ^a	1-117	13.2	P63272-1
DSIF	SPT5	1-1087	121.0	O00267-1
SPT6	SPT6 ^a	1-1726	199.0	Q7KZ85-1
	Template	48	14.7	
Nucleic acid	Non-template	48	14.9	
	RNA ^b	46	15.4	
Final	20 polypeptides, 3 nucleic acids	10,723 aa, 142 nt	1257.6	

List of all protein and nucleic acid components of EC*. For details of the complex assembly and composition, refer to the main text and Methods. aa, amino acids; nt, nucleotides; kDa, kilodalton.

^aConstruct possesses residual amino acids from the TEV or 3C protease cleavage site, respectively, that are not reported in this table. ^bBears 5'-6 FAM label and the mass of the label is included in the molecular weight.

Extended Data Table 2 | Cryo-EM data collection, refinement and validation statistics

	Map A (EMDB- 0031) (PDB 6GMH)	Map B (EMDB- 0030)	Map C (EMDB- 0032)	Map D (EMDB- 0033)	Map E (EMDB- 0034)	Map F (EMDB- 0035)	Map G (EMDB- 0036)	Map H (EMDB- 0037)
Data collection and processing								
Magnification	130,000	130,000	130,000	130,000	130,000	130,000	130,000	130,000
Voltage (kV)	300	300	300	300	300	300	300	300
Electron exposure (e-/Å ²)	34-47	34-47	34-47	34-47	34-47	34-47	34-47	34-47
Defocus range (µm)	0.4-3.5	0.4-3.5	0.4-3.5	0.4-3.5	0.4-3.5	0.4-3.5	0.4-3.5	0.4-3.5
Pixel size (Å)	1.049	1.049	1.049	1.049	1.049	1.049	1.049	1.049
Symmetry imposed	C1	C1	C1	C1	C1	C1	C1	C1
Initial particle images (no.)	1779614	1779614	1779614	1779614	1779614	1779614	1779614	1779614
Final particle images (no.)	374964	105854	69279	194782	103357	87511	313552	107349
Map resolution (Å)	3.10	3.34	3.49	3.20	3.28	3.49	3.10	3.34
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.54-9.5							
Refinement								
Initial model used (PDB code)	3PSI, 3OW8, 5OIK, 6GME							
Model resolution (Å)	3.94							
FSC threshold	0.5							
Model resolution range (Å)	2.54-9.5							
Map sharpening <i>B</i> factor (Å ²)	Locally filtered map without sharpening							
Model composition								
Non-hydrogen atoms	50240							
Protein residues	6856							
Ligands	10							
<i>B</i> factors (Å ²)								
Protein	182.20							
Ligand	164.72							
R.m.s. deviations								
Bond lengths (Å)	0.007							
Bond angles (°)	1.229							
Validation								
MolProbity score	1.64							
Clashscore	5.49							
Poor rotamers (%)	0.53							
Ramachandran plot								
Favored (%)	95.07							
Allowed (%)	4.83							
Disallowed (%)	0.10							

Extended Data Table 3 | X-ray data collection and refinement statistics SPT6 tSH2

	SPT6 tSH2
Data collection	
Space group	P 1 2 ₁ 1
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	51.02, 81.1, 51.37
α , β , γ (°)	90, 115.079, 90
Resolution (Å)	46.21 - 1.80 (1.87 - 1.80)
<i>R</i> _{sym} or <i>R</i> _{merge}	0.1034 (1.095)
<i>I</i> / σI	11.88 (0.90)
CC _{1/2}	0.997 (0.395)
Completeness (%)	94.91 (71.59)
Redundancy	6.1 (4.1)
Refinement	
Resolution (Å)	46.21 - 1.8
No. reflections	33241 (2484)
<i>R</i> _{work} / <i>R</i> _{free}	0.19 (0.32) / 0.22 (0.33)
No. atoms	3254
Protein	3071
Ligand/ion	
Water	183
<i>B</i> -factors	43.85
Protein	43.79
Ligand/ion	
Water	44.89
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.11

A single crystal was used for all data processing. Values in parentheses are for the highest-resolution shell.

The gravitationally unstable gas disk of a starburst galaxy 12 billion years ago

K. Takaki^{1*}, D. Iono^{1,2}, M. S. Yun³, I. Aretxaga⁴, B. Hatsukade⁵, D. H. Hughes⁴, S. Ikarashi⁶, T. Izumi¹, R. Kawabe^{1,2,7}, K. Kohno^{5,8}, M. Lee^{1,9}, Y. Matsuda^{1,2}, K. Nakanishi^{1,2}, T. Saito¹⁰, Y. Tamura⁹, J. Ueda¹, H. Umehata^{5,11}, G. W. Wilson³, T. Michiyama^{1,2}, M. Ando^{1,2} & P. Kamieneski³

Galaxies in the early Universe that are bright at submillimetre wavelengths (submillimetre-bright galaxies) are forming stars at a rate roughly 1,000 times higher than the Milky Way. A large fraction of the new stars form in the central kiloparsec of the galaxy^{1–3}, a region that is comparable in size to the massive, quiescent galaxies found at the peak of cosmic star-formation history⁴ and the cores of present-day giant elliptical galaxies. The physical and kinematic properties inside these compact starburst cores are poorly understood because probing them at relevant spatial scales requires extremely high angular resolution. Here we report observations with a linear resolution of 550 parsecs of gas and dust in an unlensed, submillimetre-bright galaxy at a redshift of $z = 4.3$, when the Universe was less than two billion years old. We resolve the spatial and kinematic structure of the molecular gas inside the heavily dust-obscured core and show that the underlying gas disk is clumpy and rotationally supported (that is, its rotation velocity is larger than the velocity dispersion). Our analysis of the molecular gas mass per unit area suggests that the starburst disk is gravitationally unstable, which implies that the self-gravity of the gas is stronger than the differential rotation of the disk and the internal pressure due to stellar-radiation feedback. As a result of the gravitational instability in the disk, the molecular gas would be consumed by star formation on a timescale of 100 million years, which is comparable to gas depletion times in merging starburst galaxies⁵.

Since the discovery of submillimetre-bright galaxies (SMGs) at high redshift^{6,7} two decades ago, studies of their global physical properties, such as redshift, gas mass and kinematics, have helped us to understand the origin of the extreme starburst^{8–12}. With the same goal in mind, we obtained observations of the CO $J = 4–3$ emission line in the $z = 4.3$ SMG COSMOS-AzTEC-1 (hereafter ‘AzTEC-1’) with the highest angular resolution yet achieved using the Atacama Large Millimeter/submillimetre Array (ALMA). AzTEC-1 is one of the brightest unlensed objects of this type, with an extraordinarily high star formation rate of $1,186^{+36}_{-291} M_{\odot} \text{ yr}^{-1}$ (where M_{\odot} is the mass of the Sun) and a compact starburst with a half-light radius of $R_{1/2} = 1.1 \pm 0.1$ kpc measured in the 860- μm continuum¹³. These ALMA observations resolve the CO emission at a resolution of $0.08''$ (550 pc in the physical scale) to reveal the morphology and kinematics of molecular gas within the central 2 kpc of the galaxy. In Fig. 1 we show ALMA maps of the CO line and the dust continuum at 3.2 mm and 860 μm , the velocity field and the velocity dispersion. The spatial distributions of the CO line and the 3.2-mm continuum emission independently confirm the existence of two off-centre clumps (clump 2 and clump 3), which were first detected in the 860- μm continuum¹³. Previous lower-resolution ($0.15''–0.3''$) observations^{1–3,14} have found that SMGs and optically selected massive galaxies are associated with a very compact and dusty star-forming region with $R_{1/2} = 1–2$ kpc. However, our higher-resolution

data demonstrate that the central structure of molecular gas and dust is more complicated than just a single, compact component. Such clumps of molecular gas are also seen in the central disk of the $z = 3$ gravitationally lensed star-forming galaxy SDP 81^{15,16}.

In addition, we made a $0.06''$ -resolution CO cube and a $0.05''$ -resolution 860- μm continuum image with different visibility weightings, to filter out the underlying disk emission and to highlight the clump structures (Methods). The higher-resolution velocity-integrated CO maps show that the clumps of molecular gas are aligned with the dusty star-forming clumps in the 860- μm continuum (Fig. 2). They are the second- and third-brightest clumps of 11 clumps identified previously¹³ at 860 μm . Because the brightest clump is very close to the nucleus, it is difficult to isolate even at a resolution of $0.06''$. Other faint star-forming clumps are not detected in the CO data, probably owing to poor sensitivity.

We fit the CO spectra of the clumps with a single Gaussian to derive full-width at half-maximum (FWHM) line widths of $250 \pm 50 \text{ km s}^{-1}$ and $240 \pm 50 \text{ km s}^{-1}$ for clumps 2 and 3, respectively. These line widths are one or two orders of magnitude larger than those of giant molecular clouds in nearby galaxies¹⁷. The integrated CO line flux is $S_{\text{CO}} dv = 0.056 \pm 0.009 \text{ Jy km s}^{-1}$ for clump 2 and $S_{\text{CO}} dv = 0.042 \pm 0.007 \text{ Jy km s}^{-1}$ for clump 3, indicating that each clump contains only a few per cent of the total gas mass. (Here and elsewhere, the errors quoted correspond to one standard deviation.) Adopting a CO-to- H_2 conversion factor of $\alpha_{\text{CO}} = 0.8 M_{\odot} (\text{K km s}^{-1} \text{ pc}^2)^{-1}$ and a CO excitation of $R_{41} = 0.91$, we derive gas masses of $M_{\text{CO,gas}} = (2.2 \pm 0.3) \times 10^9 M_{\odot}$ and $M_{\text{CO,gas}} = (1.7 \pm 0.3) \times 10^9 M_{\odot}$ for the two gas clumps (Methods), which are 3–5 orders of magnitude larger than the virial mass of giant molecular clouds. Therefore, these giant clumps are completely different from giant molecular clouds in nearby galaxies.

We fit the CO cube with a dynamical model to derive the kinematic properties of the CO-emitting gas. The observed velocity field is well characterized by a rotating disk with $R_{1/2} = 1.05 \pm 0.02$ kpc, a deprojected maximum rotation speed of $v_{\text{max}} = 227^{+5}_{-6} \text{ km s}^{-1}$ and a local velocity dispersion of $\sigma_0 = 74 \pm 1 \text{ km s}^{-1}$. The starburst gas disk is rotation-dominated with a ratio of rotation velocity to velocity dispersion of $v_{\text{max}}/\sigma_0 = 3.1 \pm 0.1$. In the local Universe, 80% of massive early-type galaxies with stellar masses of $\log(M_{\text{star}}/M_{\odot}) > 11.8$ exhibit dispersion-dominated stellar kinematics with $v_{\text{max}}/\sigma_0 < 1$, whereas less-massive ones are rotation-dominated^{18,19}. Given the large stellar mass of $M_{\text{star}} = (9.9^{+0.4}_{-2.6}) \times 10^{10} M_{\odot}$ (Methods), AzTEC-1 is near the massive end at $z = 4$ and might eventually evolve to become one of the most massive early-type galaxies at $z = 0$. If molecular gas and stars share the same kinematics, then the observed properties of the rotating disk suggest that the most massive galaxies do not lose much of their angular momentum during the formation phase; instead, they lose it during the subsequent evolution phase, such as during major mergers²⁰.

¹National Astronomical Observatory of Japan, Tokyo, Japan. ²SOKENDAI (The Graduate University for Advanced Studies), Tokyo, Japan. ³Department of Astronomy, University of Massachusetts, Amherst, MA, USA. ⁴Instituto Nacional de Astrofísica, Opticay Electrónica (INAOE), Puebla, Mexico. ⁵Institute of Astronomy, Graduate School of Science, The University of Tokyo, Tokyo, Japan. ⁶Kapteyn Astronomical Institute, University of Groningen, Groningen, The Netherlands. ⁷Department of Astronomy, The University of Tokyo, Tokyo, Japan. ⁸Research Center for the Early Universe, The University of Tokyo, Tokyo, Japan. ⁹Division of Particle and Astrophysical Science, Nagoya University, Nagoya, Japan. ¹⁰Max-Planck-Institute for Astronomy, Heidelberg, Germany. ¹¹RIKEN Cluster for Pioneering Research, Saitama, Japan. *e-mail: takaki.ken@nao.ac.jp

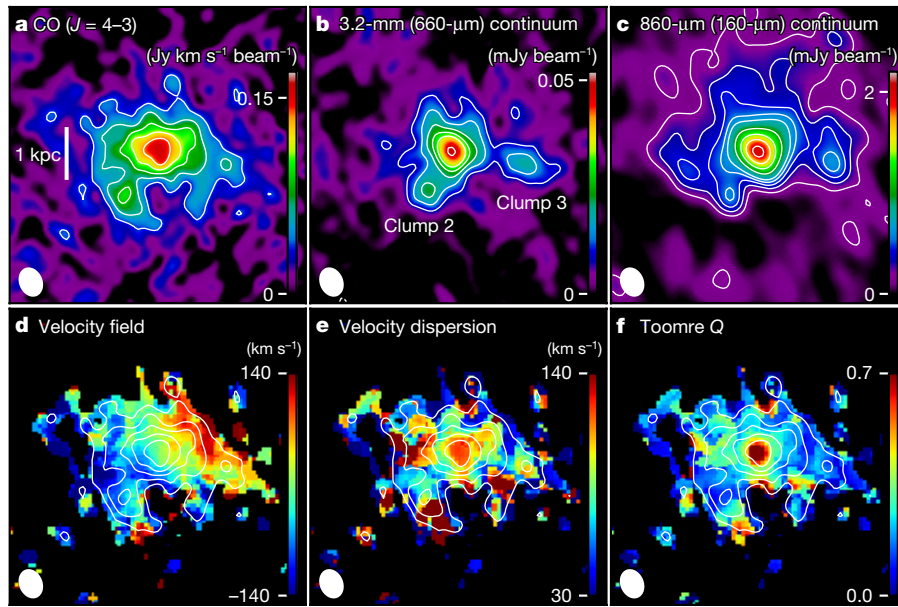


Fig. 1 | CO morphology and kinematics of AzTEC-1. **a–f**, ALMA maps of the CO ($J = 4-3$) line (**a**), 3.2-mm continuum (**b**) and 860- μm continuum (**c**), velocity field (**d**), velocity dispersion (**e**) and Toomre Q parameter (**f**). The numbers in parentheses in **b** and **c** refer to the rest-frame wavelength. The angular resolution (indicated by the white ellipses)

is $0.093'' \times 0.072''$ in all cases. The CO line is integrated in the velocity range -315 km s^{-1} to $+315 \text{ km s}^{-1}$. Contours in **a–c** are plotted every 2σ from 3σ to 11σ and every 5σ from 11σ , where 1σ is the noise level; the contours in **a** are also overplotted in **d–f**.

Until recently, clumpy rotating disks at high redshift have been discovered from observations of ionized gas²¹. Now, higher-resolution observations of molecular gas using ALMA can be used similarly. Observational and numerical studies show that giant clumps are spawned by gravitational instability in the outskirts of gas-rich disks and migrate inward by dynamical friction^{22,23}. Using the ALMA maps of the CO line intensity and velocity dispersion without any correction

for beam smearing (Fig. 1), we compute the local Toomre Q parameter, which describes the balance between self-gravity of molecular gas and turbulent pressure by stellar radiation and other sources. A thick, rotating disk can become unstable against local axisymmetric perturbations²⁴ if $Q < Q_{\text{cri}} = 0.7$. The local Q values that we measured are less than Q_{cri} over the entire disk, indicating that the gas disk should fragment and collapse through gravitational instability in the inter-clump

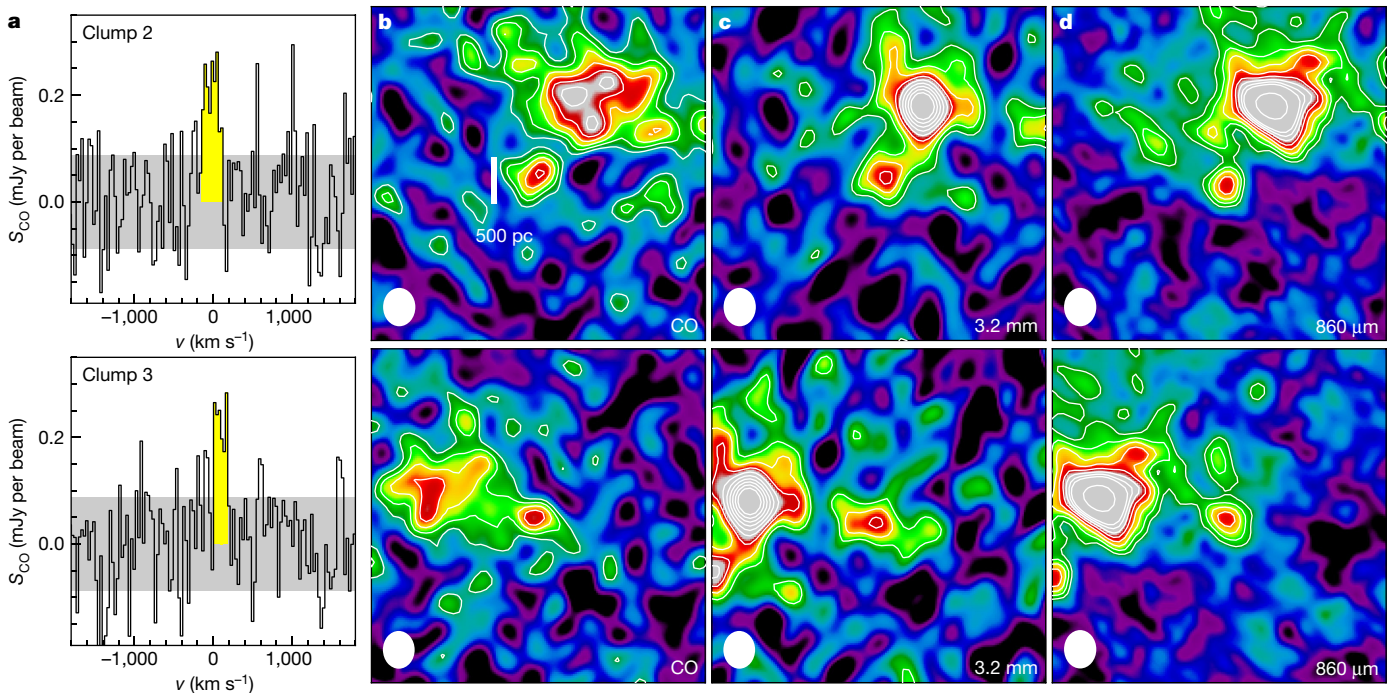


Fig. 2 | Spectra and maps of the two large clumps. **a**, For Clump 2 (top) and clump 3 (bottom), CO spectra are extracted from the Briggs-weighted cube with an angular resolution of $0.069'' \times 0.058''$. The grey shaded region indicates the standard deviation of the noise spectra. **b–d**, ALMA maps of the CO line (**b**), 3.2-mm continuum (**c**) and 860- μm continuum (**d**) for the

two clumps. The CO flux densities are integrated over the velocity range indicated by the yellow shaded regions in **a**. White filled circles represent the angular resolution of each map. The contours are plotted every 1σ from 2σ in **b** and **c** and from 4σ in **d**, and at every 5σ from 10σ .

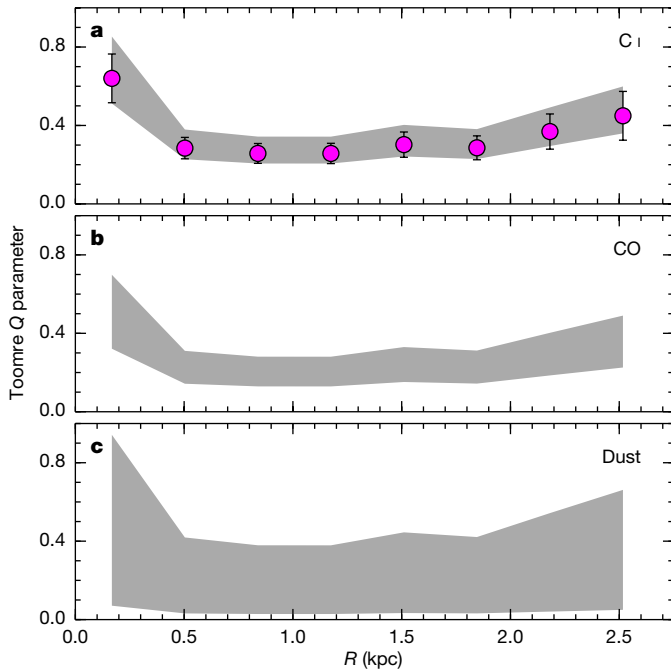


Fig. 3 | Radially averaged Toomre Q parameter. **a**, Magenta circles show the Q values computed using the C I-based gas mass. We adopt a carbon abundance relative to molecular hydrogen of $X_{\text{C I}} = 4 \times 10^{-5}$. The error bars include the uncertainties of the C I excitation temperature and the kinematic parameters as well as the measurement errors of the CO flux (Methods). The grey shaded region indicates the effect of different carbon abundances, $X_{\text{C I}} = (3\text{--}5) \times 10^{-5}$. **b**, **c**, We also compute the Q values using the CO-based (**b**) and dust-based (**c**) gas mass. The grey shaded regions illustrate the effect of changing assumptions of the ratio of the luminosities of the CO ($J = 4\text{--}3$) and CO ($J = 1\text{--}0$) lines, $R_{41} = 0.46\text{--}1.0$ (**b**), and of the dust-to-gas mass ratio, $\delta_{\text{GDR}} = 44\text{--}589$ (**c**). R is the radial distance from the centre of the galaxy.

regions. On the other hand, $Q < Q_{\text{cri}}$ at the clump locations means that the gas is gravitationally bound rather than gravitationally unstable. We also derive radially averaged Q parameters using the best-fit kinematic parameters with corrections for beam smearing and inclination. Here, the uncertainties in Q arise mainly from measurements of gas mass. We tackle this issue by determining three independent estimates of gas mass, from the C I ($J = 2\text{--}1$), CO ($J = 4\text{--}3$) and dust continuum data. All three methods indicate that $Q < 1$ is in the central 2.5 kpc of the galaxy, even after allowing for some variations in carbon abundance, CO excitation and the gas-to-dust mass ratio (Fig. 3).

In the current framework of galaxy evolution, galaxies self-regulate star formation with a marginally unstable disk^{25,26}. If a galactic disk is unstable with $Q < Q_{\text{cri}}$, intense stellar radiation temporarily boosts turbulent pressure and heats the disk until $Q > Q_{\text{cri}}$. Once the disk is stable, star formation becomes inefficient, leading to a drop in turbulent pressure. On the other hand, gas accretion may increase the gas mass per unit area in the disk, and when the increased self-gravity of the gas overcomes the decreased pressure the disk becomes unstable again. Thus, galaxy disks are kept marginally unstable with $Q \approx Q_{\text{cri}}$. In AzTEC-1, stellar radiation pressure is unlikely to support the self-gravity of gas, resulting in small Q values across the entire disk. The local velocity dispersion increases only slightly as the star-formation rate per unit area increases (Fig. 4), which suggests that stellar feedback by intense star formation does not control the velocity dispersion in the molecular gas in this case. We also find that the velocity dispersion of $\sigma \approx 100 \text{ km s}^{-1}$ in the two clumps is not much higher than in the rest of the disk. Our results imply that star-forming clumps are stable and not disrupted by radiative feedback. On the other hand, there is a strong correlation between the molecular gas mass per unit area (Σ_{gas}) and the star-formation rate per unit area (Σ_{SFR}), fitted by

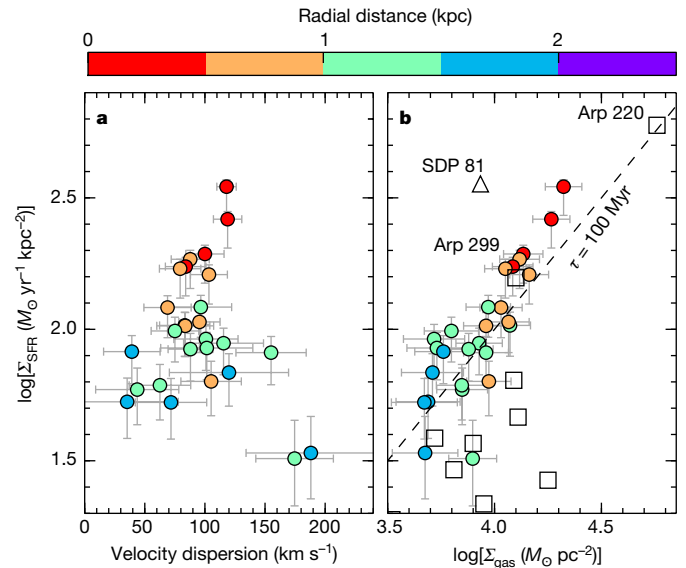


Fig. 4 | Pixel-to-pixel variations in the physical properties. **a**, **b**, The local velocity dispersion, gas mass per unit area Σ_{gas} and star-formation rate per unit area Σ_{SFR} are computed in a beam area of 0.344 kpc^2 . The colour coding indicates the radial distance from the centre of the galaxy. Open squares and an open triangle show the galaxy-averaged values in nearby starburst galaxies⁵ and the resolved value in a $z = 3$ gravitationally lensed galaxy¹⁶, respectively. The error bars reflect the overall uncertainty, including the measurement uncertainties in the velocity dispersion, the 860- μm continuum flux and the CO line flux in each pixel, and the systematic uncertainties in the total star-formation rate and the total gas mass.

the linear relation $\log[\Sigma_{\text{SFR}}/(M_{\odot} \text{ yr}^{-1} \text{ kpc}^{-2})] = (1.4 \pm 0.2) \times \log[\Sigma_{\text{gas}}/(M_{\odot} \text{ pc}^{-2})] + (-3.6 \pm 0.7)$. The gas mass per unit area derived for AzTEC-1 is extremely high, with $\log[\Sigma_{\text{gas}}/(M_{\odot} \text{ pc}^{-2})] = 3.8\text{--}4.4$, similar in magnitude to that seen in nearby starburst galaxies⁵. The implied gravitational instability is a consequence of the strong concentration of molecular gas.

In such a gravitationally unstable gas disk, molecular clouds are expected to be converted into stars efficiently. The gas depletion time in the starburst disk, defined as the gas mass divided by the star-formation rate, is comparable to the galaxy-averaged gas depletion time in nearby starburst galaxies (Fig. 4). The molecular gas reservoir of AzTEC-1 will be consumed by star formation within 100 million years, a timescale that is roughly ten times shorter than the gas depletion time in star-forming galaxies at $z = 1\text{--}3$ ²⁷ and comparable to the gas depletion times in nearby merging galaxies such as Arp 220 and Arp 299⁵ and the $z = 3$ lensed star-forming galaxy SDP 81¹⁶. An extreme starburst at high redshift may occur over a very short timescale, resulting in episodic bright periods in the submillimetre band. Otherwise, it requires new gas flowing into the central region to maintain the current level of star-formation activity.

It is still uncertain how a large amount of molecular gas is concentrated in the central 2 kpc of the galaxy. A gas-rich major merger is the most straightforward scenario, because several numerical simulations have successfully reproduced the physical properties of SMGs²⁸, including the compact gas distribution and the enhanced star-forming activity. We cannot necessarily reject the major merger scenario for rotating disk because nearby merger remnants frequently host a rotationally supported structure²⁹; however, we do not have direct evidence for a major merger in AzTEC-1. In addition to a past gas-rich major merger, multiple gas-rich minor mergers or clumpy gas streams could also lead to gas transport to the central 2 kpc³⁰. Isolated galaxies require a non-axisymmetric structure such as spiral arms or a bar to remove the angular momentum and transport a large amount of gas into the galaxy centre. AzTEC-1 does not have such a non-axisymmetric structure. To determine the roles of major mergers in extreme starbursts, we need to investigate morphological and kinematic structures in a large sample of

high-redshift SMGs using high-resolution (less than $0.1''$) and sensitive observations with ALMA.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0443-1>.

Received: 12 April 2018; Accepted: 29 June 2018;

Published online 29 August 2018.

1. Swinbank, A. M. et al. Intense star formation within resolved compact regions in a galaxy at $z = 2.3$. *Nature* **464**, 733–736 (2010).
2. Ikarashi, S. et al. Compact starbursts in $z \sim 3$ –6 submillimeter galaxies revealed by ALMA. *Astrophys. J.* **810**, 133 (2015).
3. Simpson, J. M. et al. The SCUBA-2 cosmology legacy survey: ALMA resolves the rest-frame far-infrared emission of sub-millimeter galaxies. *Astrophys. J.* **799**, 81 (2015).
4. van Dokkum, P. et al. Forming compact massive galaxies. *Astrophys. J.* **813**, 23 (2015).
5. Kennicutt, R. C. Jr. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498**, 541–552 (1998).
6. Hughes, D. H. et al. High-redshift star formation in the Hubble Deep Field revealed by a submillimetre-wavelength survey. *Nature* **394**, 241–247 (1998).
7. Barger, A. J. et al. Submillimetre-wavelength detection of dusty star-forming galaxies at high redshift. *Nature* **394**, 248–251 (1998).
8. Chapman, S. C. et al. A redshift survey of the submillimeter galaxy population. *Astrophys. J.* **622**, 772–796 (2005).
9. Bothwell, M. S. et al. A survey of molecular gas in luminous sub-millimetre galaxies. *Mon. Not. R. Astron. Soc.* **429**, 3047–3067 (2013).
10. Ivison, R. J. et al. Herschel-ATLAS: a binary HyLIRG pinpointing a cluster of starbursting protoellipticals. *Astrophys. J.* **772**, 137 (2013).
11. Tacconi, L. J. et al. Submillimeter galaxies at $z \sim 2$: evidence for major mergers and constraints on lifetimes, IMF, and CO-H₂ conversion factor. *Astrophys. J.* **680**, 246–262 (2008).
12. Hodge, J. A. et al. Evidence for a clumpy, rotating gas disk in a submillimeter galaxy at $z = 4$. *Astrophys. J.* **760**, 11 (2012).
13. Iono, D. et al. Clumpy and extended starbursts in the brightest unlensed submillimeter galaxies. *Astrophys. J.* **829**, L10 (2016).
14. Tadaki, K.-i. et al. Bulge-forming galaxies with an extended rotating disk at $z \sim 2$. *Astrophys. J.* **834**, 135 (2017).
15. Swinbank, A. M. et al. ALMA resolves the properties of star-forming regions in a dense gas disk at $z \sim 3$. *Astrophys. J.* **806**, L17 (2015).
16. Sharda, P. et al. Testing star formation laws in a starburst galaxy at redshift 3 resolved with ALMA. *Mon. Not. R. Astron. Soc.* **477**, 4380–4390 (2018).
17. Bolatto, A. D. et al. The resolved properties of extragalactic giant molecular clouds. *Astrophys. J.* **686**, 948–965 (2008).
18. Cappellari, M. Structure and kinematics of early-type galaxies from integral field spectroscopy. *Annu. Rev. Astron. Astrophys.* **54**, 597–665 (2016).
19. Veale, M. et al. The MASSIVE survey – V. Spatially resolved stellar angular momentum, velocity dispersion, and higher moments of the 41 most massive local early-type galaxies. *Mon. Not. R. Astron. Soc.* **464**, 356–384 (2017).
20. Naab, T. et al. The ATLAS^{3D} project – XXV. Two-dimensional kinematic analysis of simulated galaxies and the cosmological origin of fast and slow rotators. *Mon. Not. R. Astron. Soc.* **444**, 3357–3387 (2014).

21. Genzel, R. et al. The SINS survey of $z \sim 2$ galaxy kinematics: properties of the giant star-forming clumps. *Astrophys. J.* **733**, 101 (2011).
22. Bournaud, F. et al. The long lives of giant clumps and the birth of outflows in gas-rich galaxies at high-redshift. *Astrophys. J.* **780**, 57–75 (2014).
23. Mandelker, N. et al. The population of giant clumps in simulated high- z galaxies: in situ and ex situ migration and survival. *Mon. Not. R. Astron. Soc.* **443**, 3675–3702 (2014).
24. Genzel, R. et al. The SINS/zC-SINF survey of $z \sim 2$ galaxy kinematics: evidence for gravitational quenching. *Astrophys. J.* **785**, 75 (2014).
25. Thompson, T. et al. Radiation pressure-supported starburst disks and active galactic nucleus fueling. *Astrophys. J.* **630**, 167–185 (2005).
26. Cacciato, M. et al. Evolution of violent gravitational disc instability in galaxies: late stabilization by transition from gas to stellar dominance. *Mon. Not. R. Astron. Soc.* **421**, 818–831 (2012).
27. Tacconi, L. J. et al. Phibss: molecular gas content and scaling relations in $z \sim 1$ –3 massive, main-sequence star-forming galaxies. *Astrophys. J.* **768**, 74 (2013).
28. Narayanan, D. et al. The star-forming molecular gas in high-redshift submillimetre galaxies. *Mon. Not. R. Astron. Soc.* **400**, 1919–1935 (2009).
29. Ueda, J. et al. Cold molecular gas in merger remnants. I. Formation of molecular gas disks. *Astrophys. J. Suppl. Ser.* **214**, 1 (2014).
30. Dekel, A. et al. Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* **457**, 451–454 (2009).

Acknowledgements We thank J. Baba for discussions about a gravitational instability in SMGs. This work was supported by JSPS KAKENHI JP17J04449. We thank the ALMA staff and in particular the EA-ARC staff for their support. This research has made use of data from ALMA and HerMES project (<http://hermes.sussex.ac.uk/>). ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (South Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. HerMES is a Herschel Key Programme utilizing Guaranteed Time from the SPIRE instrument team, ESAC scientists and a mission scientist. Data analysis was in part carried out on the common-use data analysis computer system at the Astronomy Data Center (ADC) of the National Astronomical Observatory of Japan.

Reviewer information *Nature* thanks F. Bournaud and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.T. led the project and reduced the ALMA data. K.T. and D.I. wrote the manuscript. M.S.Y. reduced the Large Millimeter Telescope data and edited the final manuscript. Other authors contributed to the interpretation and commented on the ALMA proposal and the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0443-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0443-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to K.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample. AzTEC-1 was first discovered as one of the brightest sources in a 1.1-mm continuum survey of the COSMOS field obtained using the AzTEC bolometer camera on the James Clerk Maxwell Telescope (JCMT)³¹. Follow-up observations with the Redshift Search Receiver on the Large Millimeter Telescope (LMT) detected CO (4–3) and CO (5–4) lines and yielded a spectroscopic redshift of $z = 4.342$, which is also confirmed by the detection of the [C I] line using the Submillimetre Array³². Previous ALMA observations of the 860- μm continuum emission at 0.02'' resolution revealed that AzTEC-1 is composed of a compact core, an extended disk and multiple 200-pc clumps within the disk¹³. The half-light radius for the 860- μm continuum emission is $R_{1/2} = 1.1 \pm 0.1$ kpc. The rest-frame ultraviolet continuum emission is not resolved even by Hubble Space Telescope (HST)/WFC3 imaging, which is also suggestive of compact emission with $R_{1/2} < 2.6$ kpc³³. In this work, the Chabrier initial mass function³⁴ and the following cosmological parameters are assumed: present-day Hubble parameter $H_0 = 70$ km s⁻¹ Mpc⁻¹, matter-density parameter $\Omega_M = 0.3$ and density parameter for the cosmological constant $\Omega_\Lambda = 0.7$. At $z = 4.342$, an angular scale of 0.1'' corresponds to a physical scale of 670 pc.

Observations. In AzTEC-1, we carried out observations of the CO (4–3) emission line at the rest-frame frequency of 461.040 GHz (86.309 GHz in the observed frame), with ALMA band-3 receivers covering the frequency range of 85.4–89.1 GHz and 97.5–101.2 GHz in two array configurations and baseline lengths of 41 m to 16.2 km. The shortest 5th-percentile baseline is 600 m, corresponding to the maximum recoverable scale of 1.15'' at 86.3 GHz. The observations were executed on 2017 October (C43-10 array configuration) and November (C43-8). On-source time was 5.8 h and 1.2 h, respectively. The total observing time including calibration and overhead was 14 h. We used the Common Astronomy Software Application (CASA) package³⁵ for data calibration. We first estimated the continuum flux density in the frequency range excluding 86.1–86.5 GHz, and then subtracted it from the data in the visibility plane using the CASA/*uvcontsub* task. We used the CASA/*tlean* task with natural weighting to make a cube with a velocity width of 30 km s⁻¹. The resultant angular resolution and the noise level are $0.093'' \times 0.072''$ (624 pc \times 483 pc in physical scale) and $1\sigma = 78$ μJy per 30 km s⁻¹, respectively. We cleaned down to the 2σ noise level in a circular mask with a radius of 0.4''. We also made a high-resolution cube of the CO (4–3) line with a Briggs weighting (robust parameter of +0.5). The angular resolution is then $0.069'' \times 0.058''$ (470 pc \times 390 pc) and the noise level is $1\sigma = 87$ μJy per 30 km s⁻¹. To show the significance of the clump detection in AzTEC-1, we also made a $0.055'' \times 0.042''$ map of the 860- μm continuum emission using archival ALMA data¹³. We adopted a taper of 0.03'' in the visibility plane, resulting in a standard deviation of 47 μJy . We use the high-resolution cube and map only for studying the clump properties in Fig. 2.

Extended Data Fig. 1 shows the galaxy-integrated CO (4–3) spectrum extracted within an aperture of 0.8'' centred on AzTEC-1. The Gaussian line width derived is 305 ± 17 km s⁻¹ (FWHM). We made the CO moment maps of velocity-integrated intensity, velocity field and velocity dispersion in the velocity range between -315 km s⁻¹ and $+315$ km s⁻¹ using the CASA/*immoments* task. A 2σ masking threshold was adopted when creating the velocity field and velocity dispersion maps. The total CO line flux measured is $S_{\text{CO}} dv = 1.84 \pm 0.17$ Jy km s⁻¹ with 0.8'' aperture photometry in the velocity-integrated intensity map. The uncertainty in the CO line flux measurement is estimated by placing 300 random apertures in the same map. The CO line flux and the velocity width measured by the LMT are³² $S_{\text{CO}} dv = 1.75 \pm 0.24$ Jy km s⁻¹ and $\Delta v = 380$ km s⁻¹, in good agreement with the ALMA-derived flux of $S_{\text{CO}} dv = 1.60 \pm 0.13$ Jy km s⁻¹ with $\Delta v = 390$ km s⁻¹.

We also created two 3.2-mm line-free continuum maps with the same angular resolution as the CO (4–3) cubes by excluding the CO frequency range. The root-mean-square noise is 3.0 μJy per beam in the $0.093'' \times 0.072''$ map and 3.3 μJy per beam in the $0.069'' \times 0.058''$ map. We derived a total flux density of $S_{3.2\text{mm}} = 273 \pm 41$ μJy with an aperture of 0.8'', consistent with the 3-mm continuum flux density of $S_{3\text{mm}} = 300 \pm 40$ μJy from Plateau de Bure interferometer observations with a 6'' beam³⁶.

We made follow-up observations of the C I (1–0) line at 92.134 GHz in the observed frame and the C I (2–1) line at 151.511 GHz with ALMA band-3 and band-4 receivers in March 2018. We reduced the data in a similar way as for the CO (4–3) data and created a cube with a spectral resolution of 30 km s⁻¹ and a map of the 2.1-mm continuum with a Briggs weighting (robust parameter of +0.5). The angular resolution is $1.7'' \times 1.1''$ in the C I (1–0) map and $0.8'' \times 0.7''$ in the C I (2–1) map. The noise level is $1\sigma = 0.49$ mJy in each 30 km s⁻¹ channel in the C I (1–0) cube, $1\sigma = 0.38$ mJy in the C I (2–1) cube and $1\sigma = 20$ μJy in the 2.1-mm continuum map. For the flux measurements of the two C I lines, we integrated the line flux density in the same velocity range as for the CO (4–3) line. We also made a natural weighted C I (2–1) map with the same angular resolution as for the C I (1–0) map using the CASA/*imsmooth* task, which is used to obtain a C I (1–0)/C I

(2–1) line ratio. In Extended Data Fig. 1 and Extended Data Table 1, we show the line spectra and tabulate the measured line fluxes and luminosities. The 2.1-mm continuum flux density is $S_{2.1\text{mm}} = 989 \pm 20$ μJy . We also detected the CO (7–6) emission line at 151.007 GHz, but do not discuss this information here.

Global SED properties of AzTEC-1. We collected the photometric data for AzTEC-1 from the latest multi-wavelength catalogues (Subaru³⁷, VISTA³⁷, Spitzer³⁷, Herschel^{38,39} and VLA⁴⁰). After excluding marginal detections below 5σ and adding our ALMA photometry at 860 μm , 2.1 mm and 3.2 mm, we constrained the global spectral energy distribution (SED) from optical to radio (Extended Data Fig. 2). To account for possible zero-point offsets, we added a systematic uncertainty of 0.1 mag to the flux errors in the optical and near-infrared bands. Using the MAGPHYS code^{41,42}, we fitted the observed SED to stellar population synthesis models⁴³, taking into account dust attenuation and dust emission in a physically consistent way. The best-fitting SED model indicates that AzTEC-1 is a massive, star-forming galaxy with a stellar mass of $M_{\text{star}} = (9.9^{+0.4}_{-2.6}) \times 10^{10} M_\odot$ and a star-formation rate of $\text{SFR} = 1,186^{+36}_{-291} M_\odot \text{yr}^{-1}$. The dust emission is characterized by a total infrared luminosity of $L_{\text{dust}} = (1.9^{+0.0}_{-0.3}) \times 10^{13} L_\odot$, a dust mass of $M_{\text{dust}} = (1.1 \pm 0.2) \times 10^9 M_\odot$ and a dust temperature of $T_{\text{dust}} = 43^{+4}_{-2}$ K. The uncertainties are based on the 2.5th–97.5th-percentile range of the probability distributions.

Gas mass. We derived two independent estimates of the molecular gas mass for AzTEC-1 using the CO (4–3) line and [C I] line luminosities. For the gas-mass estimates based on the CO (4–3) luminosity, there are uncertainties about the CO excitation $R_{41} = L'_{\text{CO}(4-3)}/L'_{\text{CO}(1-0)}$ and the CO-to-H₂ conversion factor $\alpha_{\text{CO}} = M_{\text{gas}}/L_{\text{CO}}$. Alternatively, the [C I] line is an independent, optically thin tracer of cold molecular gas mass in nearby and high-redshift galaxies^{44–48}, and having both [C I] line measurements provides useful constraints on the physical conditions of the emitting gas. First, we estimated an excitation temperature of $T_{\text{ex}} = 27.7 \pm 4.8$ K from the C I (1–0)/C I (2–1) line ratio of $R_{\text{CI}} = 0.52 \pm 0.13$ in the $1.7'' \times 1.1''$ resolution maps, using the relation⁴⁵ $T_{\text{ex}} = 38.8 \text{ K} / \ln(2.11/R_{\text{CI}})$. Then, using the C I (2–1) line flux in the $0.8'' \times 0.7''$ map, we computed a neutral carbon mass of $M_{\text{CI}} = (1.7 \pm 0.3) \times 10^7 M_\odot$ from $M_{\text{CI}} = 4.566 \times 10^{-4} Q(T_{\text{ex}}) \times 1/5 \times \exp(62.5/T_{\text{ex}}) L'_{\text{CI}(2-1)}$, where $Q(T_{\text{ex}}) = 1 + 3\exp(-23.6/T_{\text{ex}}) + 5\exp(-62.5/T_{\text{ex}})$ is the partition function⁴⁵. The uncertainty in the neutral carbon mass includes the error in the flux measurement and the uncertainty in the excitation temperature. The molecular gas mass derived from the [C I] line luminosity is $M_{\text{CI,gas}} = (7.2 \pm 1.3) \times 10^{10} M_\odot$, adopting carbon abundance of $X_{\text{CI}} = 4 \times 10^{-5}$, which is the average of the typical value of 3×10^{-5} in normal star-forming galaxies^{44–48} and the elevated value of 5×10^{-5} in the central region of the local starburst galaxy M82⁴⁹. The resulting gas-to-dust mass ratio $M_{\text{CI,gas}}/M_{\text{dust}} = 65 \pm 17$ is smaller than the average value of $\delta_{\text{GDR}} = 120 \pm 28$ in 18 nearby starburst galaxies⁵⁰, but is still in the 5th–95th-percentile range, $\delta_{\text{GDR}} = 44–589$. The CO (4–3) luminosity also gives a gas mass of $M_{\text{CO,gas}} = (6.6 \pm 0.6) \times 10^{10} \times \alpha_{\text{CO}}/0.8 \times (1.0/R_{41}) M_\odot$. If the CO (4–3) line is thermalized ($R_{41} = 1$) and the conversion factor of $\alpha_{\text{CO}} = 0.8 M_\odot (\text{K km s}^{-1} \text{pc}^2)^{-1}$ is used^{11,12,51,52}, then the CO-based gas mass is similar to the C I-based gas mass. For consistency between C I and CO, we adopt a gas excitation of $R_{41} = 0.91$, which is larger than the average value for SMGs at high redshift ($R_{41} = 0.46$) and comparable with the average value for quasi-stellar objects ($R_{41} = 0.87$)⁵³. Adopting $\alpha_{\text{CO}} = 4 M_\odot (\text{K km s}^{-1} \text{pc}^2)^{-1}$, commonly used in normal star-forming galaxies⁵¹, is not appropriate because the CO-based gas mass substantially exceeds the C I-based gas mass. Modelling suggests that CO emission in dense clumps is more highly excited, compared with the entire disk⁵⁴. If the gas is thermalized with $R_{41} = 1$, then the gas mass of clumps can be 10% smaller.

SFR and gas mass per unit area. We obtained the total star-formation rate $\text{SFR}_{\text{total}}$ and gas mass $M_{\text{gas,total}}$ in AzTEC-1 as mentioned above. Because the 860- μm (160 μm in the rest frame) continuum flux density $S_{860\mu\text{m}}$ traces star formation, we compute star-formation-rate surface densities in each pixel as $\Sigma_{\text{SFR}} = \text{SFR}_{\text{total}} \times (S_{860\mu\text{m}}/S_{860\mu\text{m,total}})/\Omega_{\text{beam}}$, where $S_{860\mu\text{m,total}} = 16.9 \pm 0.7$ mJy and Ω_{beam} is the effective beam area of 0.344 kpc². Here, we use the 860- μm continuum map with a pixel scale of 0.07'' to avoid oversampling, although the original pixel scale is 0.01''. The uncertainties of Σ_{SFR} include errors in the flux measurements of $S_{860\mu\text{m}}$ and $S_{860\mu\text{m,total}}$ and systematic errors in the SED modelling. In a similar way, using the CO (4–3) map, we derive gas mass surface densities as $\Sigma_{\text{gas}} = M_{\text{gas,total}} \times (S_{\text{CO}} dv / S_{\text{CO,total}} dv) / \Omega_{\text{beam}}$.

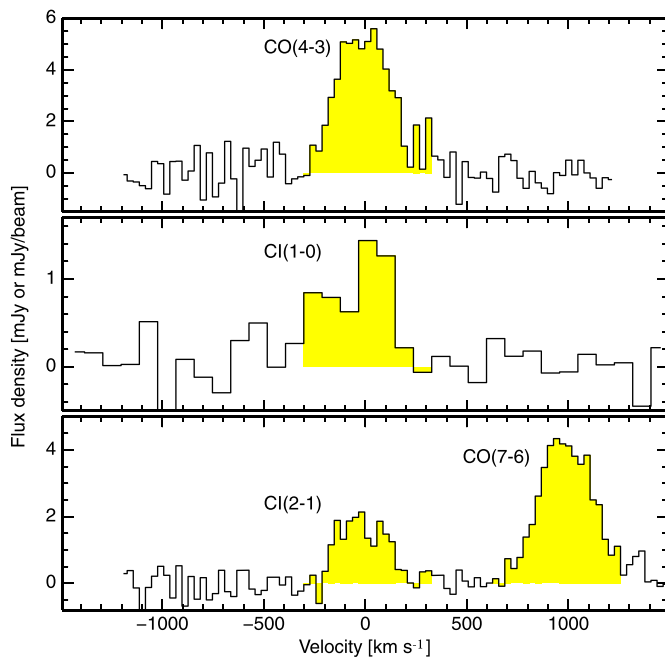
Disk modelling and dynamical mass. We fit the natural-weighted CO cube with dynamical models of a disk galaxy using the GalPaK^{3D} code⁵⁵. We adopt a thick exponential disk with an arctan rotation curve of $v(R) \propto v_{\text{max}} \arctan(R/R_t)$, where v_{max} is the maximum circular velocity and R_t is the turnover radius. A model galaxy consists of ten free parameters: centroid position (x, y), systematic velocity v_{sys} , line flux $S dv$, half-light radius $R_{1/2}$, turnover radius R_t , inclination i , position angle PA, maximum circular velocity v_{max} and velocity dispersion σ_0 . These parameters are convolved with the clean beam and are fitted to the data cube using a Markov chain Monte Carlo (MCMC) algorithm. The CO spectra extracted along

the kinematic major axis in the observed cube together with the best-fitting model are shown in Extended Data Fig. 3. The observed CO kinematics is well characterized by a rotating disk. The best-fit values are $SdV = 1.88_{+0.02-0.01} \text{ Jy km s}^{-1}$, $R_{1/2} = 1.05 \pm 0.02 \text{ kpc}$, $R_t = 0.18 \pm 0.03 \text{ kpc}$, $i = 44^\circ \pm 1^\circ$, $PA = -64^\circ \pm 1^\circ$, $v_{\text{max}} = 227_{-6}^{+5} \text{ km s}^{-1}$ and $\sigma_0 = 74 \pm 1 \text{ km s}^{-1}$. We adopt the median and the 95% confidence interval of the last 60% of the MCMC chain for 20,000 iterations as the best-fit values and the uncertainties (Extended Data Fig. 4). For a symmetric oblate disk, the inclination corresponds to the projected minor-to-major-axis ratio of $q_{\text{obs}} = 0.73 \text{ asin}^2(i) = (1 - q_{\text{obs}}^2)/(1 - q_{\text{int}}^2)$, assuming a disk thickness of $q_{\text{int}} = 0.15$. **Toomre Q parameter.** In a thin rotating gas disk with epicyclic frequency κ , the dispersion relation for an axisymmetric perturbation is $\omega^2 = \kappa^2 - 2\pi G \Sigma_{\text{gas}} |k| + \sigma_{0,\text{gas}}^2 k^2$, where ω is the growth rate and k is the wavenumber of the perturbation^{56–58}. The perturbation grows exponentially in time when $\omega^2 < 0$, leading to gravitational collapse of gas clouds. This condition is characterized by the Toomre Q parameter $Q = \kappa \sigma_{0,\text{gas}} / (\pi G \Sigma_{\text{gas}})$, and the threshold value is $Q_{\text{cri}} = 1$ for a thin gas disk and $Q_{\text{cri}} = 0.67$ for a thick disk^{21,26}. When the disk consists of two components (gas and stars) with the same velocity dispersion, the threshold value increases to³⁹ $Q_{\text{cri},2\text{com}} = 1.3$. The self-gravity of gas overcomes the repelling forces by pressure and differential rotation when $Q < Q_{\text{cri}}$. Using the maximum circular velocity derived from the disk model and the measured molecular gas mass per unit area and velocity dispersion without correction for beam-smearing, we estimated the local Q parameter in each pixel assuming a flat rotation curve with⁵⁸ $\kappa = 1.4v_{\text{max}}/R$ (Fig. 2). In Fig. 3, we show radially averaged CO fluxes and Q parameters along elliptical rings with the axis ratio of 0.73 using the best-fit kinematic parameters with correction for beam smearing and inclination.

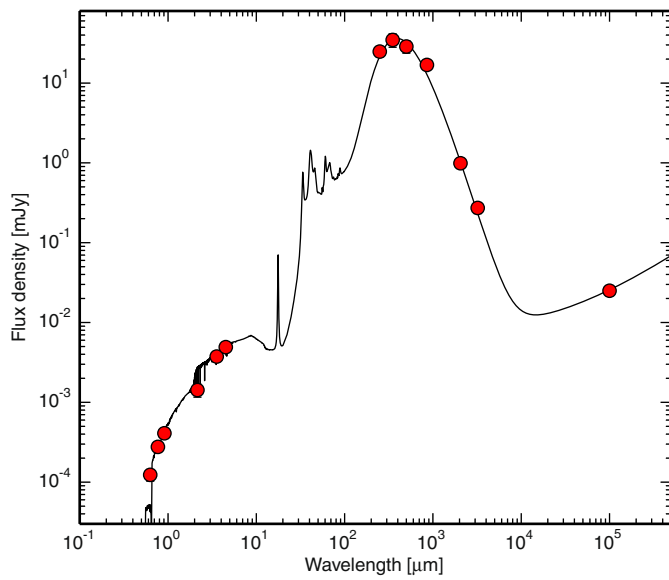
Code availability. The ALMA data were reduced using the CASA pipeline version 5.1.1, available at https://casa.nrao.edu/casa_obtaining.shtml. The disk modelling code GalPaK^{3D} is publicly available at <http://galpak.irap.omp.eu55>.

Data availability. This work makes use of the following ALMA data: ADS/JAO.ALMA#2017.1.00300.S and 2017.A.00032.S. Calibrated data that support the findings of this study are publicly available in the ALMA archive (https://almascience.eso.org/aq/?project_code=2017.1.00300.S, https://almascience.eso.org/aq/?project_code=2017.A.00032.S). The HerMES data were obtained through the Herschel Database in Marseille (HeDaM; <http://hedam.lam.fr>), which is operated by CeSAM and hosted by the Laboratoire d'Astrophysique de Marseille.

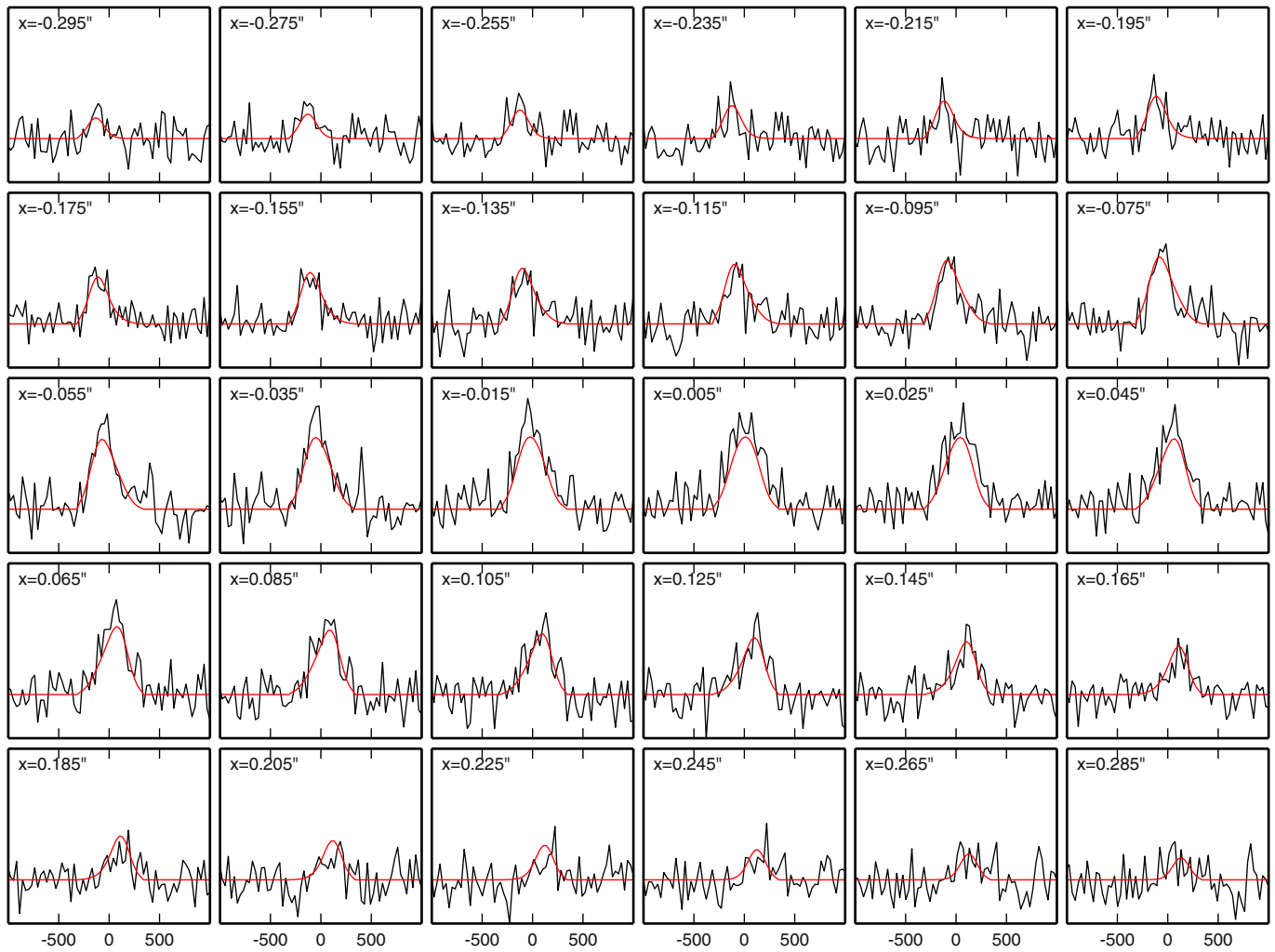
31. Scott, K. B. et al. AzTEC millimetre survey of the COSMOS field – I. Data reduction and source catalogue. *Mon. Not. R. Astron. Soc.* **385**, 2225–2238 (2008).
32. Yun, M. S. et al. Early science with the Large Millimeter Telescope: CO and [C II] emission in the $z = 4.3$ AzTEC J095942.9+022938 (COSMOS AzTEC-1). *Mon. Not. R. Astron. Soc.* **454**, 3485–3499 (2015).
33. Toft, S. et al. Submillimeter galaxies as progenitors of compact quiescent galaxies. *Astrophys. J.* **782**, 68 (2014).
34. Chabrier, G. The galactic disk mass function: reconciliation of the Hubble Space Telescope and nearby determinations. *Astrophys. J.* **586**, L133–L136 (2003).
35. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. CASA architecture and applications. *ASP Conf. Ser.* **376**, 127–130 (2007).
36. Smolčić, V. et al. The redshift and nature of AzTEC/COSMOS 1: a starburst galaxy at $z = 4.6$. *Astrophys. J.* **731**, L27 (2011).
37. Laigle, C. et al. The COSMOS2015 catalog: exploring the $1 < z < 6$ universe with half a million galaxies. *Astrophys. J. Suppl. Ser.* **24**, 224 (2016).
38. Roseboom, I. G. et al. The Herschel multi-tiered extragalactic survey: SPIRE-mm photometric redshifts. *Mon. Not. R. Astron. Soc.* **419**, 2758–2773 (2012).
39. Oliver, S. J. et al. The Herschel multi-tiered extragalactic survey: HerMES. *Mon. Not. R. Astron. Soc.* **424**, 1614–1635 (2012).
40. Smolčić, V. et al. The VLA-COSMOS 3 GHz large project: continuum data and source catalog release. *Astron. Astrophys.* **602**, A1 (2017).
41. da Cunha, E., Charlot, S. & Elbaz, D. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Mon. Not. R. Astron. Soc.* **388**, 1595–1617 (2008).
42. da Cunha, E. et al. An ALMA survey of sub-millimeter galaxies in the extended Chandra deep field south: physical properties derived from ultraviolet-to-radio modeling. *Astrophys. J.* **806**, 110 (2015).
43. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344**, 1000–1028 (2003).
44. Papadopoulos, P. P., Thi, W.-F. & Viti, S. C I lines as tracers of molecular gas, and their prospects at high redshifts. *Mon. Not. R. Astron. Soc.* **351**, 147–160 (2004).
45. Weiß, A. et al. Gas and dust in the Cloverleaf quasar at redshift 2.5. *Astron. Astrophys.* **409**, L41–L45 (2003).
46. Weiß, A. et al. Atomic carbon at redshift ~ 2.5 . *Astron. Astrophys.* **429**, L25–L28 (2005).
47. Danielson, A. L. R. et al. The properties of the interstellar medium within a star-forming galaxy at $z = 2.3$. *Mon. Not. R. Astron. Soc.* **410**, 1687–1702 (2011).
48. Bothwell, M. S. et al. ALMA observations of atomic carbon in $z \sim 4$ dusty star-forming galaxies. *Mon. Not. R. Astron. Soc.* **466**, 2825–2841 (2017).
49. White, G. J. et al. CO and C I maps of the starburst galaxy M 82. *Astron. Astrophys.* **284**, L23–L26 (1994).
50. Wilson, C. et al. Luminous infrared galaxies with the submillimeter array. I. Survey overview and the central gas to dust ratio. *Astrophys. J. Suppl. Ser.* **178**, 189–224 (2008).
51. Bolatto, A. D., Wolfire, M. & Leroy, A. K. The CO-to-H₂ conversion factor. *Annu. Rev. Astron. Astrophys.* **51**, 207–268 (2013).
52. Downes, D. & Solomon, P. M. Rotating nuclear rings and extreme starbursts in ultraluminous galaxies. *Astrophys. J.* **507**, 615–654 (1998).
53. Carilli, C. L. & Walter, F. Cool gas in high-redshift galaxies. *Annu. Rev. Astron. Astrophys.* **51**, 105–161 (2013).
54. Bournaud, F. et al. Modeling CO emission from hydrodynamic simulations of nearby spirals, starbursting mergers, and high-redshift galaxies. *Astron. Astrophys.* **575**, A56 (2015).
55. Bouche, N. et al. GalPak3D: a Bayesian parametric tool for extracting morphokinematics of galaxies from 3D data. *Astrophys. J.* **150**, 92 (2015).
56. Toomre, A. On the gravitational stability of a disk of stars. *Astrophys. J.* **139**, 1217–1238 (1964).
57. Wang, B. et al. Gravitational instability and disk star formation. *Astrophys. J.* **427**, 759–769 (1994).
58. Binney, J. & Tremaine, S. *Galactic Dynamics* 2nd edn, 494–496 (Princeton Univ. Press, Princeton, 2008).
59. Romeo, A. B. & Wiegert, J. The effective stability parameter for two-component galactic discs: is $Q^{-1} \approx Q_{\text{stars}}^{-1} + Q_{\text{gas}}^{-1}$? *Mon. Not. R. Astron. Soc.* **416**, 1191–1196 (2011).



Extended Data Fig. 1 | Galaxy-integrated CO (4–3), CO (1–0) and C I (2–1) spectra of AzTEC-1. The CO (4–3) spectrum is extracted using an $0.8''$ -diameter aperture in the natural-weighted map cube. The C I (1–0) and C I (2–1) spectra are extracted from the peak positions in map cubes with $1.7'' \times 1.1''$ and $0.8'' \times 0.7''$ resolution, respectively. Yellow shaded regions show the velocity range $v = -315 \text{ km s}^{-1}$ to $v = +315 \text{ km s}^{-1}$, in which the velocity-integrated line fluxes are measured.

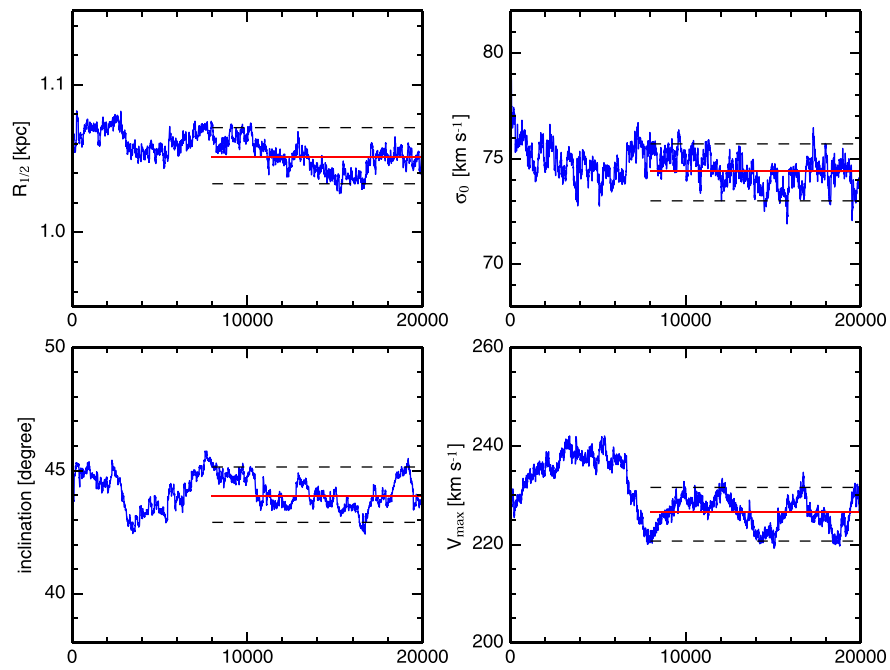


Extended Data Fig. 2 | Galaxy-integrated SED of AzTEC-1. Red circles show the photometric data from Subaru (r' , i' , z')³⁷, VISTA (K_s)³⁷, Spitzer (3.6 μm , 4.4 μm)³⁷, Herschel (250 μm , 350 μm , 500 μm)^{38,39}, ALMA (860 μm , 2.1 mm, 3.2 mm) and JVLA (10 cm)⁴⁰. The black line shows the best-fitting SED model from MAGPHYS^{41,42}.



Extended Data Fig. 3 | CO spectra along the kinematic major axis. Spectra are extracted at a position angle of $PA = -64^\circ$. The spatial offset x from the galactic centre is shown at the upper left of each panel. Red

lines indicate the spectra of the best-fitting dynamical model produced by GalPaK^{3D}.



Extended Data Fig. 4 | Full MCMC chain for 20,000 iterations. Red solid lines and black dashed lines indicate the median and 95% confidence interval of the last 60% of the MCMC chain.

Extended Data Table 1 | Line fluxes in AzTEC-1

Line	Frequency (GHz)	$S_{\text{line}} dv$ (Jy km s ⁻¹)	L'_{line} (10 ¹⁰ K km s ⁻¹ pc ²)
CO (4-3)	461.041	1.84±0.17*	8.21±0.78
Cl (1-0)	492.161	0.45±0.08†	1.76±0.30
Cl (2-1)	809.342	0.49±0.09‡	0.70±0.13
Cl (2-1)	809.342	0.63±0.11†	0.92±0.16

*The flux within a 0.8" aperture in the 0.093" × 0.072" map.

†The peak flux in the 1.7" × 1.1" map.

‡The peak flux in the 0.8" × 0.7" map.

Probing high-momentum protons and neutrons in neutron-rich nuclei

The CLAS Collaboration*

The atomic nucleus is one of the densest and most complex quantum-mechanical systems in nature. Nuclei account for nearly all the mass of the visible Universe. The properties of individual nucleons (protons and neutrons) in nuclei can be probed by scattering a high-energy particle from the nucleus and detecting this particle after it scatters, often also detecting an additional knocked-out proton. Analysis of electron- and proton-scattering experiments suggests that some nucleons in nuclei form close-proximity neutron–proton pairs^{1–12} with high nucleon momentum, greater than the nuclear Fermi momentum. However, how excess neutrons in neutron-rich nuclei form such close-proximity pairs remains unclear. Here we measure protons and, for the first time, neutrons knocked out of medium-to-heavy nuclei by high-energy electrons and show that the fraction of high-momentum protons increases markedly with the neutron excess in the nucleus, whereas the fraction of high-momentum neutrons decreases slightly. This effect is surprising because in the classical nuclear shell model, protons and neutrons obey Fermi statistics, have little correlation and mostly fill independent energy shells. These high-momentum nucleons in neutron-rich nuclei are important for understanding nuclear parton distribution functions (the partial momentum distribution of the constituents of the nucleon) and changes in the quark

distributions of nucleons bound in nuclei (the EMC effect)^{1,13,14}. They are also relevant for the interpretation of neutrino-oscillation measurements¹⁵ and understanding of neutron-rich systems such as neutron stars^{3,16}.

Since the 1950s, the independent-particle shell model has been an indispensable guide for understanding nuclei¹⁷. In this model, nucleons move independently in well defined quantum orbits (shells) with low momentum, k ($k < k_F$, where k_F is the Fermi momentum), similarly to electrons in atoms. The potential in which the nucleons move is the average nuclear field created by their mutual strong interactions. Although successful in making many important predictions, such as shell closures and the spins and parities of nuclear ground and excited states, this textbook picture of the nucleus is incomplete: electron-scattering experiments in nuclei ranging from lithium to lead measured only about 60%–70% of the expected number of protons in each shell¹⁸. Newer shell-model-type calculations include the effects of long-range correlations, increasing this fraction to about 80%¹⁹.

Modern superconducting accelerators—with high energy, high intensity and high duty factor—enable experiments that use scattering reactions to resolve the structure and dynamics of individual nucleons and nucleon pairs in nuclei. The resolving power of a measurement is determined by its momentum transfer, and its interpretation relies on

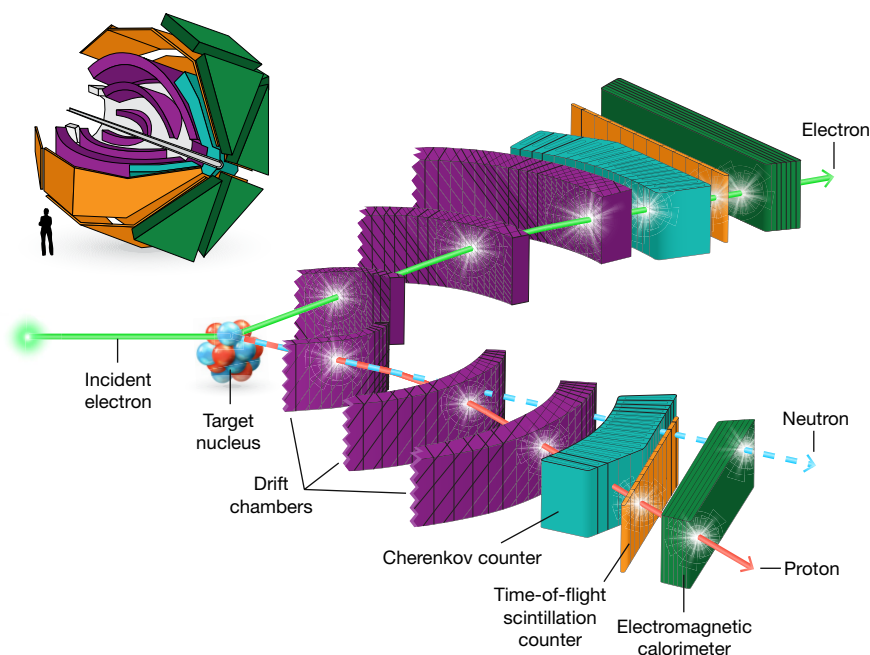


Fig. 1 | CLAS spectrometer. Two segments of the CLAS spectrometer. Electrons travelling with energies of up to 6 GeV hit nuclei, knocking out individual protons and neutrons. The momenta of the scattered electrons and knocked-out protons are reconstructed by analysing their trajectories as they bend in a toroidal magnetic field. The neutron

momenta are deduced from their time of flight until they interact with the electromagnetic calorimeter. Inset, the almost-spherical CLAS. The electron beam travels along the grey pipe, hitting a target near the centre of the spectrometer.

*A list of authors and their affiliations appears at the end of the paper.

the theoretical modelling of the interaction, which should account for all possible mechanisms that lead to the same measured final state. The high-momentum transfer measurements reported here are discussed in terms of interaction with single nucleons, which is the simplest reaction picture that is consistent with the measured observables^{1–3} and various *ab initio* calculations²⁰.

When analysed within this framework, electron-scattering experiments suggest that about 20% of the nucleons in nuclei have momentum greater than k_F ^{1–3,10–12}. These nucleons are absent in the one-body shell-model description of the data and are coupled into short-lived correlated nucleon pairs with large relative momentum ($k_{\text{relative}} > k_F \approx 250 \text{ MeV } c^{-1}$) and small centre-of-mass momentum (k_{CM}), referred to as short-range correlated (SRC) pairs^{1–3}.

The dominant force between nucleons in SRC pairs is tensor in nature^{1,2}. This pair-wise interaction acts predominantly on spin-1 neutron–proton (*np*) SRC pairs, leading to a predominance of *np* SRC pairs over proton–proton (*pp*) and neutron–neutron (*nn*) SRC pairs by a factor of about 20. This phenomenon is referred to as ‘*np* dominance’^{1–8}.

Almost all high-momentum nucleons in nuclei belong to an SRC pair. As the short-distance interaction between nucleons in SRC pairs is very strong, the characteristics of the resulting pairs are largely independent of the rest of the nucleus. Thus, the distribution of high-momentum nucleons (the ‘high-momentum tail’ of the distribution) has a universal shape for all nuclei^{1–3,9–11,21}.

SRC pairs considerably complicate the nuclear ground state and nuclear-structure calculations. From a theoretical point of view, one can use a unitary transform to shift this complexity from the ground state to many-body interaction operators that describe the same measured final state^{22,23}, shifting the physics from high-momentum correlations to short-distance operators. The physical pictures of high-momentum nucleons and short-distance operators are based on the different momentum and distance scales of these effects from those of the shell model. The results reported here constrain short-distance phenomena, as described in either framework.

The analysis reported here was motivated by the quest to study the short-distance dynamics of protons and neutrons in neutron-rich nuclei. For the first time, we simultaneously measured electron-induced quasi-elastic knock-out of protons and neutrons from medium and heavy nuclei, using the $A(e,e'p)$ and $A(e,e'n)$ reactions, respectively (*e*, incident electron; *e'*, scattered electron; *A*, target nucleus). The simultaneous measurement of both proton and neutron knock-out allows us to directly compare their properties using minimal assumptions. Analysed within the one-body reaction picture, the data from these measurements perform four functions: (1) quantifying the relative fractions of high-momentum ($k > k_F$) protons and neutrons, (2) showing that adding neutrons to the nucleus increases the fraction of high-momentum protons, (3) helping confirm the *np* dominance of the high-momentum tail in medium and heavy nuclei, and (4) supporting momentum-sharing inversion in heavy nuclei. In a more general framework, the data show that short-distance dynamics is similar in all nuclei, supporting a scale separation of short-distance physics from the nuclear shell model.

The data presented here were collected in 2004 in Hall-B of the Thomas Jefferson National Accelerator Facility (Jefferson Laboratory) in Virginia, USA, and were reanalysed as part of the data-mining initiative of the Jefferson Laboratory. The experiment used a 5.014 GeV electron beam incident on deuterium, carbon, aluminium, iron and lead targets, and the CEBAF (continuous electron beam accelerator facility) large acceptance spectrometer (CLAS)²⁴ to detect the scattered electrons and any associated hadrons knocked out during the interaction (see Fig. 1). CLAS used a toroidal magnetic field and six independent sets of drift chambers, time-of-flight scintillation counters, Cherenkov counters and electromagnetic calorimeters, covering scattering angles from about 8° to 140°, for charged-particle identification and trajectory reconstruction. The neutrons were identified by observing interactions in the forward electromagnetic calorimeters (covering about 8°–45°) with no associated charged-particle tracks in the drift

chambers. The angle- and momentum-dependent neutron detection efficiency and momentum reconstruction resolution were measured simultaneously using the $d(e,e'p\pi^+\pi^-n)$ reaction (*d*, deuterium; π , pion; see Supplementary Information). The experiment recorded all events with a scattered electron detected in both the electromagnetic calorimeter and the Cherenkov counter, along with any other particles.

High-energy electrons scatter from the nucleus by transferring a single virtual photon, carrying momentum q and energy ω . In quasi-elastic scattering, this momentum transfer is absorbed by a nucleon with initial momentum p_i . If the nucleon does not rescatter as it leaves the nucleus, then it will emerge with final momentum $p_f = p_i + q$. Thus, we can reconstruct the approximate initial momentum of the nucleon from the missing momentum, namely, the difference between the detected final momentum and the transferred momentum: $p_{\text{miss}} = p_f - q$. Similarly, the excitation energy of the residual ($A-1$) nucleus is related to the missing energy, $E_{\text{miss}} = \omega - T_f$, where T_f is the nucleon’s kinetic energy.

Although this quasi-elastic picture of the scattering reaction is highly intuitive and consistent with the measured observables, other reaction mechanisms using two-body currents that result in the same measured final state are added coherently and cannot be distinguished from the quasi-elastic mechanism. Contribution from non-quasi-elastic reaction mechanisms is minimized by the use of large momentum transfer and the specific reaction kinematics used in the measurement (see Methods). In addition, these effects are further diminished by forming cross-section ratios.

In this analysis, we studied $(e,e'p)$ and $(e,e'n)$ quasi-elastic knock-out event samples measured in two kinematical regions, corresponding to electron scattering from high-initial-momentum ($p_i > k_F$) nucleons, presumably from an SRC pair, or from low-initial-momentum ($p_i < k_F$) nucleons, presumably from shell-model states.

Using these events, we derived both the ratio of $A(e,e'n)/A(e,e'p)$ events for each region and the double ratio of high-momentum (SRC) to low-momentum (shell model) nucleons in nuclei relative to carbon $[A(e,e'x)_{\text{high}}/A(e,e'x)_{\text{low}}]/[{}^{12}\text{C}(e,e'x)_{\text{high}}/{}^{12}\text{C}(e,e'x)_{\text{low}}]$, where *A* stands for Al, Fe or Pb. The double ratio is simply an estimator for the

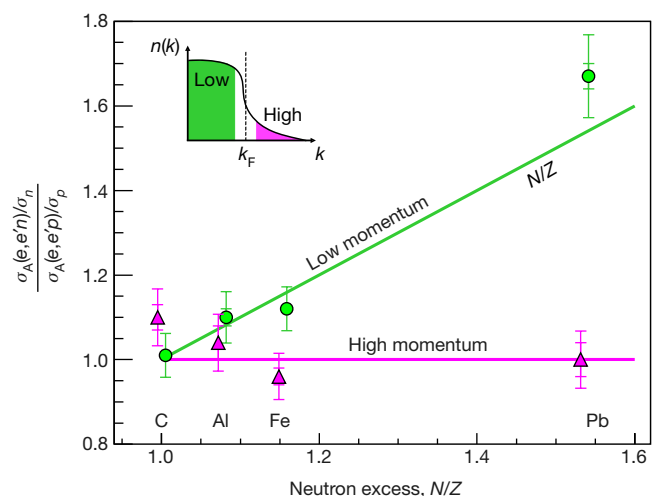


Fig. 2 | Relative abundances of high- and low-initial-momentum neutrons and protons. Reduced cross-section ratio, $[\sigma_A(e,e'n)/\sigma_n]/[\sigma_A(e,e'p)/\sigma_p]$, for low-momentum (green circles) and high-momentum (purple triangles) events. The inset illustrates a typical nuclear momentum distribution as a function of nucleon momentum, where ‘low’ and ‘high’ refer to the initial nucleon momentum. The lines show the simple N/Z behaviour (green), as expected from the number of neutrons and protons in the nucleus for low-momentum nucleons, and the prediction of the *np*-dominance model (purple; $[\sigma_A(e,e'n)/\sigma_n]/[\sigma_A(e,e'p)/\sigma_p] = 1$) for high-momentum nucleons. The inner error bars correspond to statistical uncertainties and the outer ones include both statistical and systematic uncertainties, both at the 1σ or 68% confidence level (see Supplementary Information).

increased fraction of SRC nucleons in an asymmetric nucleus compared to carbon. We used carbon as a reference because it is a well studied, medium-mass symmetric nucleus and has similar average density to the other nuclei measured here. In addition, forming cross-section ratios relative to carbon significantly reduces the effects of detector acceptance and efficiency corrections (see Supplementary Information). For each kinematical setting, we used the same selection criteria on the detected scattered electron and associated knocked-out nucleon to select quasi-elastic $A(e,e'p)$ and $A(e,e'n)$ events.

Low-initial-momentum events are characterized by low missing energy and low missing momentum ($E_{\text{miss}} < 80$ –90 MeV and $p_{\text{miss}} = |\mathbf{p}_{\text{miss}}| < 250$ MeV c^{-1} , where c is the speed of light in vacuum; see Supplementary Information). Because the neutron momentum resolution was not good enough to select these events directly, we developed a set of alternative constraints to select the same events by using the detected electron momentum and the knocked-out nucleon angle, which were unaffected by the neutron momentum resolution (see Methods).

Similarly, we selected the high-initial-momentum events in two steps. We first selected quasi-elastic events with a leading nucleon by setting conditions on the energy and momentum transfer and requiring that the outgoing nucleon be emitted with most of the transferred momentum in the general direction of the momentum transfer. We then selected high-initial-momentum events by requiring large missing momentum ($p_{\text{miss}} > 300$ MeV c^{-1}). These selection criteria ensured that the electron interacted with a single high-initial-momentum proton or neutron in the nucleus^{2,3,12}. Lastly, we optimized the nucleon-momentum-dependent conditions to account for the neutron momentum reconstruction resolution and corrected for any remaining bin-migration effects (see Methods).

To verify the neutron detection efficiency, detector acceptance corrections and event selection method, we extracted the neutron-to-proton reduced cross-section ratio for carbon, for both high and low initial nucleon momenta: $[\sigma_{\text{IC}}(e,e'n)/\sigma_n]/[\sigma_{\text{IC}}(e,e'p)/\sigma_p]$ (that is, the ratio of measured cross-sections for the scattering of electrons from carbon, scaled by the known elastic-scattering electron–neutron, σ_n , and electron–proton, σ_p , cross-sections). Figure 2 shows that these two measured cross-section ratios are consistent with unity, as expected for a symmetric nucleus. This shows that in both high- and low-initial-momentum kinematics, we have restricted the reaction mechanisms to primarily quasi-elastic scattering and have correctly accounted for the various detector-related effects.

For the other measured nuclei, the low-momentum $(e,e'n)/(e,e'p)$ reduced cross-section ratios grow approximately as N/Z , as expected from the number of neutrons (N) and protons (Z) in the nucleus. However, the high-momentum $(e,e'n)/(e,e'p)$ ratios are consistent with unity for all measured nuclei (see Fig. 2).

The struck nucleons could reinteract as they emerge from the nucleus, which we refer to as final-state interaction. Such an effect would cause the number of detected outgoing nucleons to decrease and also modify the angles and momenta of the knocked-out nucleons. These effects were estimated for symmetric and asymmetric nuclei using a relativistic Glauber framework, which showed that the decrease in the measured cross-section is similar for protons and neutrons and thus has a minor impact on cross-section ratios (see Methods).

Because rescattering changes the event kinematics, some of the events with high measured p_{miss} could have originated from electron scattering from a low-initial-momentum nucleon, which then rescattered, thus increasing p_{miss} . If the high-initial-momentum (high- p_{miss}) nucleons originated from electron scattering from the more numerous low-initial-momentum nucleons, followed by nucleon rescattering, then the high-momentum $(e,e'n)/(e,e'p)$ ratio would show the same N/Z dependence as the low-momentum ratio. Because the high-momentum $(e,e'n)/(e,e'p)$ ratio is independent of A , these nucleon-rescattering effects must be small in this measurement.

Thus, the constant $(e,e'n)/(e,e'p)$ high-momentum ratios indicate that there are equal numbers of high-initial-momentum protons and

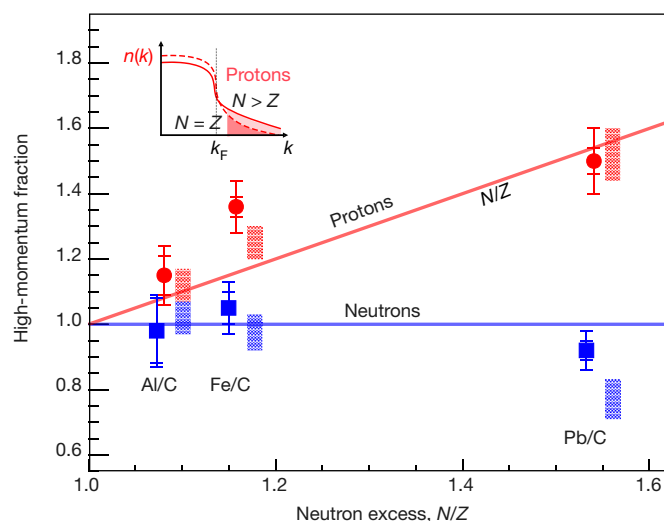


Fig. 3 | Relative high-momentum fractions for neutrons and protons.

Red circles with error bars denote the double ratio of the number of $(e,e'p)$ high-momentum proton events to low-momentum proton events for nucleus A relative to carbon. The inner error bars are statistical and the outer ones include both statistical and systematic uncertainties, both at the 1σ or 68% confidence level. Blue squares with error bars show the same for neutron events. Red and blue rectangles show the range of predictions of the phenomenological np -dominance model for proton and neutron ratios, respectively (see Supplementary Information). The red line (high-momentum fraction equal to N/Z) and the blue line (high-momentum fraction equal to 1) are drawn to guide the eye. The inset demonstrates how adding neutrons to the target nucleus (solid red curve) increases the fraction of protons in the high-momentum tail (shaded region).

neutrons in asymmetric nuclei, even though these nuclei contain up to 50% more neutrons than protons. This observation is consistent with high-initial-momentum nucleons belonging primarily to np SRC pairs, even in neutron-rich nuclei²⁵. This equality implies a greater fraction of high-initial-momentum protons. For example, if 20% of the 208 nucleons in ^{208}Pb have high initial momentum, then these consist of 21 protons and 21 neutrons. This corresponds to a high-momentum proton fraction of $21/82 \approx 25\%$ and a corresponding neutron fraction of only $21/126 \approx 17\%$.

To quantify the relative fraction of high-momentum protons and neutrons in different nuclei with minimal experimental and theoretical uncertainties, we extracted the double ratio of $(e,e'x)$ high-initial-momentum to low-initial-momentum events for nucleus A relative to carbon for both protons and neutrons. We found that the fraction of high-initial-momentum protons increases by about 50% from carbon to lead (see Fig. 3).

Moreover, the corresponding fraction of high-initial-momentum neutrons seems to decrease by about $10\% \pm 5\%$ (1σ). Nucleon rescattering, if substantial, should increase in larger nuclei and should affect protons and neutrons equally (see Methods). Because, unlike the proton ratio, the neutron ratio decreases slightly with mass number, this also rules out sizeable nucleon rescattering effects.

Figure 3 also shows the results of a simple phenomenological (that is, experiment-based) np -dominance model^{5,26} that uses a mean-field momentum distribution at low momentum ($k < k_F$) and a scaled deuteron-like high-momentum tail. This model agrees with our data and also predicts momentum-sharing inversion, that is, on average protons move faster than neutrons in neutron-rich nuclei.

These results indicate that high-momentum nucleons and short-range two-body currents are universal and independent of the shell model. This conclusion holds for both the quasi-elastic and unitary-transformed pictures of the interaction and indicate that nuclei must be viewed in a scale-dependent way: nuclear structure at higher momentum scales and shorter distances must be described

using universal two-body physics, which is absent in an independent-particle shell-model picture using one-body operators.

The surprising fact that increasing the number of neutrons in a nucleus increases the fraction of high-initial-momentum protons (proposed in ref. ²⁶ and bolstered by exact calculations of light nuclei²⁵ and calculations of heavier nuclei²⁷ and asymmetric nuclear matter²⁸) has several broad implications. Neutron stars contain about 5%–10% protons and electrons in their central layers. Our results imply that the extreme neutron excess in a neutron star could dramatically increase the effects of short-distance currents on the protons, which could affect the cooling rate and equation of state of neutron stars^{3,16}.

There is evidence that the high-momentum nucleons associated with SRC pairs are responsible for the EMC effect, that is, the change in the quark distribution of nucleons bound in nuclei^{1,13}. The EMC effect (named after the European Muon Collaboration) may result from two-body short-distance interactions that can be viewed as temporary high-density fluctuations of nucleon pairs in the nucleus, in which the internal structure of the affected nucleons is briefly modified¹. If this mechanism indeed occurs, then the average proton in neutron-rich nuclei (the minority species) is more likely to belong to a correlated pair and should therefore be more modified than the average neutron (the majority species). Observing such increased modification of the proton structure in neutron-rich nuclei could shed new light on the currently unknown origin of these modifications of nuclear parton distribution functions.

Furthermore, the *np* dominance of SRC pairs and two-body short-distance currents in heavy nuclei has considerable implications in many areas of nuclear and particle physics (including nuclear correlation functions and the double-beta decay rate of nuclei²⁹, the nature of the repulsive core of the nucleon–nucleon interaction^{2,6} and neutrino–nucleus interactions), where high-precision extraction of oscillation parameters and searches for new physics beyond the standard model require detailed understanding of the nuclear ground state and neutrino–interaction operators¹⁵.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0400-z>.

Received: 18 December 2017; Accepted: 13 June 2018;

Published online 13 August 2018.

- Hen, O., Miller, G. A., Piasetzky, E. & Weinstein, L. B. Nucleon–nucleon correlations, short-lived excitations, and the quarks within. *Rev. Mod. Phys.* **89**, 045002 (2017).
- degli Atti, C. C. In-medium short-range dynamics of nucleons: recent theoretical and experimental advances. *Phys. Rep.* **590**, 1–85 (2015).
- Frankfurt, L., Sargsian, M. & Strikman, M. Recent observation of short-range nucleon correlations in nuclei and their implications for the structure of nuclei and neutron stars. *Int. J. Mod. Phys. A* **23**, 2991–3055 (2008).
- Subedi, R. et al. Probing cold dense nuclear matter. *Science* **320**, 1476–1478 (2008).
- CLAS Collaboration. Momentum sharing in imbalanced Fermi systems. *Science* **346**, 614–617 (2014).
- Korover, I. et al. Probing the repulsive core of the nucleon–nucleon interaction via the $^4\text{He}(e,e'pN)$ triple-coincidence reaction. *Phys. Rev. Lett.* **113**, 022501 (2014).
- Piasetzky, E., Sargsian, M., Frankfurt, L., Strikman, M. & Watson, J. W. Evidence for strong dominance of proton–neutron correlations in nuclei. *Phys. Rev. Lett.* **97**, 162504 (2006).
- Tang, A. et al. *n*–*p* short-range correlations from $(p, 2p + n)$ measurements. *Phys. Rev. Lett.* **90**, 042301 (2003).
- Fomin, N. et al. New measurements of high-momentum nucleons and short-range structures in nuclei. *Phys. Rev. Lett.* **108**, 092502 (2012).
- CLAS Collaboration. Measurement of two- and three-nucleon short-range correlation probabilities in nuclei. *Phys. Rev. Lett.* **96**, 082501 (2006).
- Frankfurt, L. L., Strikman, M. I., Day, D. B. & Sargsyan, M. Evidence for short-range correlations from high Q^2 (e,e') reactions. *Phys. Rev. C* **48**, 2451 (1993).
- Arrington, J., Higinbotham, D. W., Rosner, G. & Sargsian, M. Hard probes of short-range nucleon–nucleon correlations. *Prog. Part. Nucl. Phys.* **67**, 898–938 (2012).

- Weinstein, L. B., Piasetzky, E., Higinbotham, D. W., Gomez, J., Hen, O. & Shneor, R. Short range correlations and the EMC effect. *Phys. Rev. Lett.* **106**, 052301 (2011).
- Hen, O., Piasetzky, E. & Weinstein, L. B. New data strengthen the connection between short range correlations and the EMC effect. *Phys. Rev. C* **85**, 047301 (2012).
- Gallagher, H., Garvey, G. & Zeller, G. P. Neutrino–nucleus interactions. *Annu. Rev. Nucl. Part. Sci.* **61**, 355–378 (2011).
- Li, B. A., Cai, B. J., Chen, L. W. & Xu, J. Nucleon effective masses in neutron-rich matter. *Prog. Part. Nucl. Phys.* **99**, 29–119 (2018).
- Caurier, E., Martínez-Pinedo, G., Nowacki, F., Poves, A. & Zuker, A. P. The shell model as a unified view of nuclear structure. *Rev. Mod. Phys.* **77**, 427–488 (2005).
- Kelly, J. J. Nucleon knockout by intermediate-energy electrons. *Adv. Nucl. Phys.* **23**, 75–294 (1996).
- Dickhoff, W. H. & Barbieri, C. Self-consistent Green's function method for nuclei and nuclear matter. *Prog. Part. Nucl. Phys.* **52**, 377–496 (2004).
- Carlson, J. et al. Quantum Monte Carlo methods for nuclear physics. *Rev. Mod. Phys.* **87**, 1067–1118 (2015).
- Frankfurt, L. L. & Strikman, M. I. High-energy phenomena, short-range nuclear structure and QCD. *Phys. Rep.* **76**, 215–347 (1981).
- Bogner, S. K. & Roscher, D. High-momentum tails from low-momentum effective theories. *Phys. Rev. C* **86**, 064304 (2012).
- More, S. N., Bogner, S. K. & Furnstahl, R. J. Scale dependence of deuteron electrodisintegration. *Phys. Rev. C* **96**, 054004 (2017).
- Mecking, B. A. et al. The CEBAF large acceptance spectrometer (CLAS). *Nucl. Instrum. Methods A* **503**, 513–553 (2003).
- Wiringa, R. B., Schiavilla, R., Pieper, S. C. & Carlson, J. Nucleon and nucleon-pair momentum distributions in $A \leq 12$ nuclei. *Phys. Rev. C* **89**, 024305 (2014).
- Sargsian, M. M. New properties of the high-momentum distribution of nucleons in asymmetric nuclei. *Phys. Rev. C* **89**, 034305 (2014).
- Ryckebusch, J., Vanhalst, M. & Cosyn, W. Stylized features of single-nucleon momentum distributions. *J. Phys. G* **42**, 055104 (2015).
- Rios, A., Polls, A. & Dickhoff, W. H. Depletion of the nuclear Fermi sea. *Phys. Rev. C* **79**, 064308 (2009).
- Kortelainen, M. and Suhonen, J. Nuclear matrix elements of $0\nu\beta\beta$ decay with improved short-range correlations. *Phys. Rev. C* **76**, 024315 (2007).

Acknowledgements This work was supported by the US Department of Energy (DOE), contract number DEAC05-06OR23177, under which Jefferson Science Associates, LLC, operates the Thomas Jefferson National Accelerator Facility; by the National Science Foundation, the Israel Science Foundation; the Chilean Comisión Nacional de Investigación Científica y Tecnológica; the French Centre National de la Recherche Scientifique and Commissariat à l'Energie Atomique; the French–American Cultural Exchange; the Italian Istituto Nazionale di Fisica Nucleare; the National Research Foundation of Korea; and the UK Science and Technology Facilities Council.

Reviewer information Nature thanks T. Aumann, D. Phillips and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The CEBAF large acceptance spectrometer was designed and constructed by the CLAS Collaboration and Jefferson Laboratory. Data processing and calibration, Monte Carlo simulations of the detector and data analyses were performed by a large number of CLAS Collaboration members, who also discussed and approved the scientific results. The analysis presented here was performed by M. Duer with input from O. Hen, E. Piasetzky and L. B. Weinstein and reviewed by the CLAS Collaboration.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0400-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The CLAS Collaboration

M. Duer¹, O. Hen², E. Piasetzky¹, H. Hakobyan³, L. B. Weinstein⁴, M. Braverman¹, E. O. Cohen¹, D. Higinbotham⁵, K. P. Adhikari⁶, S. Adhikari⁷, M. J. Amarian⁴, J. Arrington⁸, A. Ashkenazi², J. Ball⁹, I. Balossino¹⁰, L. Barion¹⁰, M. Battaglieri¹¹, V. Batourine^{5,12}, A. Beck², I. Bedlinskiy¹³, A. S. Biselli^{14,15}, S. Boiarinov⁵, W. J. Briscoe¹⁶, W. K. Brooks^{3,5}, S. Bueltmann⁴, D. Bulumulla⁴, V. D. Burkert⁵, F. Cao¹⁷, D. S. Carman⁵, A. Celentano¹¹, G. Charles⁴, T. Chetry¹⁸, G. Ciullo^{10,19}, L. Clark²⁰, B. A. Clary¹⁷, P. L. Cole^{5,21,22}, M. Contalbrigo¹⁰, O. Cortes²¹, V. Crede²³, R. Cruz-Torres², A. D'Angelo^{24,25}, N. Dashyan²⁶, R. De Vita¹¹, E. De Sanctis²⁷, M. Defurne⁹, A. Deur⁵, C. Djalali²⁸, G. Dodge⁴, R. Dupre²⁹, H. Egiyan⁵, A. El Alaoui³, L. El Fassi⁶, P. Eugenio²³, R. Fersch^{30,31}, A. Filippi³², T. A. Forest²¹, G. Gavalian^{5,33}, Y. Ghandilyan²⁶, S. Gilad², G. P. Gilfoyle³⁴, K. L. Giovanetti³⁵, F. X. Girod⁵, E. Golovatch³⁶, R. W. Gothe²⁸, K. A. Griffioen³¹, L. Guo^{5,7}, N. Harrison⁵, M. Hattawy⁸, F. Hauenstein⁴, K. Hafidi⁸, K. Hicks¹⁸, M. Holtrop³³, C. E. Hyde⁴, Y. Ilieva^{16,28}, D. G. Ireland²⁰,

B. S. Ishkhanov³⁶, E. L. Isupov³⁶, K. Joo¹⁷, M. L. Kabir⁶, D. Keller³⁷, G. Khachatryan²⁶, M. Khachatryan⁴, M. Khandaker³⁸, A. Kim¹⁷, W. Kim¹², A. Klein⁴, F. J. Klein²², I. Korover¹, S. E. Kuhn⁴, L. Lanza²⁴, G. Laskaris², P. Lenisa¹⁰, K. Livingston²⁰, I. J. D. MacGregor²⁰, C. Marchand⁹, N. Markov¹⁷, B. McKinnon²⁰, S. Mey-Tal Beck², T. Mineeva³, M. Mirazita²⁷, V. Moiseev^{5,36}, R. A. Montgomery²⁰, A. Movsisyan¹⁰, C. Munoz-Camacho²⁹, B. Mustapha⁸, S. Nadeeshani⁴, P. Nadel-Turonski⁵, S. Niccolai²⁹, G. Niculescu³⁵, M. Osipenko¹¹, A. I. Ostrovidov²³, M. Paolone³⁹, E. Pasyuk⁵, M. Patsyuk², A. Papadopoulou², K. Park^{5,12}, D. Payette⁴, W. Phelps⁷, O. Pogorelko¹³, J. Poudel⁴, J. W. Price⁴⁰, S. Procureur⁹, Y. Prok^{4,37}, D. Protopopescu²⁰, M. Ripani¹¹, A. Rizzo^{24,25}, G. Rosner²⁰, P. Rossi^{5,27}, F. Sabatié⁹, A. Schmidt², C. Salgado³⁸, B. A. Schmookler², R. A. Schumacher¹⁵, E. P. Segarra², Y. G. Sharabian⁵, G. D. Smith⁴¹, D. Sokhan²⁰, N. Sparveris³⁹, S. Stepanyan⁵, S. Strauch^{16,28}, M. Taiuti⁴², J. A. Tan¹², M. Ungaro^{5,43}, H. Voskanyan²⁶, E. Voutier²⁹, D. P. Watts⁴¹, X. Wei⁵, N. Zachariou⁴¹, J. Zhang³⁷, X. Zheng^{8,37} & Z. W. Zhao⁴

¹Tel Aviv University, Tel Aviv, Israel. ²Massachusetts Institute of Technology, Cambridge, MA, USA. ³Universidad Técnica Federico Santa María, Valparaíso, Chile. ⁴Old Dominion University, Norfolk, VA, USA. ⁵Thomas Jefferson National Accelerator Facility, Newport News, VA, USA.

⁶Mississippi State University, Mississippi State, MS, USA. ⁷Florida International University, Miami, FL, USA. ⁸Argonne National Laboratory, Argonne, IL, USA. ⁹IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France. ¹⁰INFN, Sezione di Ferrara, Ferrara, Italy. ¹¹INFN, Sezione di Genova, Genova, Italy. ¹²Kyungpook National University, Daegu, South Korea. ¹³Institute of Theoretical and Experimental Physics, Moscow, Russia. ¹⁴Fairfield University, Fairfield, CT, USA. ¹⁵Carnegie Mellon University, Pittsburgh, PA, USA. ¹⁶The George Washington University, Washington, DC, USA. ¹⁷University of Connecticut, Storrs, CT, USA. ¹⁸Ohio University, Athens, OH, USA. ¹⁹Università di Ferrara, Ferrara, Italy. ²⁰University of Glasgow, Glasgow, UK. ²¹Idaho State University, Pocatello, ID, USA. ²²Catholic University of America, Washington, DC, USA. ²³Florida State University, Tallahassee, FL, USA. ²⁴INFN, Sezione di Roma Tor Vergata, Rome, Italy. ²⁵Università di Roma Tor Vergata, Rome, Italy. ²⁶Yerevan Physics Institute, Yerevan, Armenia. ²⁷INFN, Laboratori Nazionali di Frascati, Frascati, Italy. ²⁸University of South Carolina, Columbia, SC, USA. ²⁹Institut de Physique Nucléaire, CNRS/IN2P3 and Université Paris Sud, Orsay, France. ³⁰Christopher Newport University, Newport News, VA, USA. ³¹College of William and Mary, Williamsburg, VA, USA. ³²INFN, Sezione di Torino, Torino, Italy. ³³University of New Hampshire, Durham, NH, USA. ³⁴University of Richmond, Richmond, VA, USA. ³⁵James Madison University, Harrisonburg, VA, USA. ³⁶Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia. ³⁷University of Virginia, Charlottesville, VA, USA. ³⁸Norfolk State University, Norfolk, VA, USA. ³⁹Temple University, Philadelphia, PA, USA. ⁴⁰California State University, Carson, CA, USA. ⁴¹Edinburgh University, Edinburgh, UK. ⁴²Università di Genova, Genova, Italy. ⁴³Rensselaer Polytechnic Institute, Troy, NY, USA.

*e-mail: hen@mit.edu

METHODS

Analysis details. The $A(e,e'p)$ and $A(e,e'n)$ event samples were selected by determining the common angular region for detecting both protons and neutrons, correcting for their detection efficiencies and accounting for the different momentum resolutions. Neutron momenta were determined to an uncertainty of about 10%–15% from their time of flight, which was measured using the CLAS electromagnetic calorimeter. Proton momenta were determined to an uncertainty of about 1% from the curvature of their trajectories in the CLAS magnetic field.

We accounted for the momentum resolution difference by: (1) selecting the desired $A(e,e'p)$ events in high- and low-momentum kinematics, (2) 'smearing' the proton momentum for each event using the measured neutron momentum resolution, and (3) using unsmear and smeared $A(e,e'p)$ event samples to study bin migration effects and optimize the event selection criteria. This process results in a smeared event sample with as many of the 'original' $A(e,e'p)$ events as possible (that is, high selection efficiency) and as few other events as possible (that is, high purity). We used the smeared proton momenta in the final selection of $A(e,e'p)$ events for consistency with the $A(e,e'n)$ analysis.

The final event sample contains about 85%–90% of the desired sample with about 15% contamination, resulting in less than about 5% more events in our sample. This 5% cross-section correction caused a correction of less than 1% to the ratios between different nuclei. We assigned systematic uncertainties equal to these corrections. See Supplementary Information for additional details.

Non-quasi-elastic reaction mechanisms and data interpretation. Because the measurement detects only the final-state particles, we need to include different reaction mechanisms to infer information about the initial nuclear state. In addition to quasi-elastic electron scattering from a single nucleon, the full reaction mechanism could include contributions from meson-exchange currents, isobar currents (exciting the struck nucleon to an excited state) and elastic and inelastic nucleon rescattering (final-state interactions, FSIs). In the case of high missing momentum, elastic FSIs include rescattering between the nucleons of the pair or with the residual system. The relative contribution of these reaction mechanisms, as compared to the quasi-elastic reaction of interest, strongly depends on the reaction kinematics^{2,3,12,30,31}. Minimizing non-quasi-elastic reaction mechanisms also reduces their interference with the quasi-elastic reaction.

The high-missing-momentum measurement reported here was carried out using largely anti-parallel kinematics with high Q^2 and $x_B > 1$ (Q^2 , four-momentum transfer squared; x_B , Bjorken variable). This kinematical region minimizes non-quasi-elastic reaction mechanisms as follows^{2,3,12,31}: (i) meson-exchange currents are suppressed by a factor of $1/Q^2$ compared to SRC pair breakup, and their contribution in our kinematics is small; and (ii) isobar currents are suppressed at $x_B > 1$, where the virtual photon transfers less energy and is less able to excite the nucleon to an isobar current for a given Q^2 . Further, at large knock-out nucleon momenta, FSI effects can be calculated using a generalized eikonal approximation in a Glauber framework^{12,31–34}. These calculations show that in our measurement,

elastic FSIs are largely suppressed for mean-field knock-out. For SRC breakup, they are confined to nucleons in close proximity, and thus the largest part of the scattering cross-section can be attributed to SRC pairs^{12,30}.

This simple quasi-elastic picture, with suppressed FSIs, is strongly supported by the fact that it describes well both high- Q^2 electron-scattering data and high-energy proton data^{7,8}, which have very different reaction mechanisms. In addition, the results of the electron and proton-scattering experiments give consistent SRC-pair isospin ratios^{4,7,8} and centre-of-mass momentum distributions^{8,35}.

Asymmetry dependence of reaction mechanisms. As protons and neutrons propagate through asymmetric nuclei, they encounter more neutrons than protons, which could lead to different FSI effects that do not cancel in the cross-section ratios. However, at the large Q^2 of this measurement, the pp and nn scattering cross-sections are almost identical, leading to a 1% difference between proton and neutron FSIs, as estimated quantitatively using a full relativistic multiple-scattering Glauber approximation³¹.

Data interpretation using unitary transformations. From a theoretical standpoint, one can describe the scattering reaction in one of two mathematically equivalent ways: (a) using one-body operators acting on a ground-state wavefunction with a high-momentum tail, as discussed in the main text, or (b) using unitary-transformed many-body operators acting on a 'mean-field' ground state without a considerable high-momentum tail²³. In the latter case, the description of the ground state is simpler, but complicated many-body operators are needed to describe the electron–nucleus interaction that leads to the measured final state. Although this approach has been proven to be very efficient in describing long-distance and low-energy properties of nuclei, it is not clear yet if it is a cost-effective way to describe the measured short-distance physics in heavy nuclei. Therefore, we discuss our results predominantly in the framework of untransformed wavefunctions and interactions.

Data availability. The raw data from this experiment are archived in Jefferson Laboratory's mass-storage silo (<https://scicomp.jlab.org/docs/node/9>).

30. Frankfurt, L. L., Sargsian, M. M. & Strikman, M. I. Feynman graphs and generalized eikonal approach to high energy knock-out processes. *Phys. Rev. C* **56**, 1124 (1997).
31. Colle, C., Cosyn, W. & Ryckebusch, J. Final-state interactions in two-nucleon knockout reactions. *Phys. Rev. C* **93**, 034608 (2016).
32. Colle, C. et al. Extracting the mass dependence and quantum numbers of short-range correlated pairs from $A(e,e'p)$ and $A(e, e'pp)$ scattering. *Phys. Rev. C* **92**, 024604 (2015).
33. Dutta, D., Hafidi, K. & Strikman, M. Color transparency: past, present and future. *Prog. Part. Nucl. Phys.* **69**, 1–27 (2013).
34. CLAS Collaboration. Measurement of transparency ratios for protons from short-range correlated pairs. *Phys. Lett. B* **722**, 63–68 (2013).
35. Shneur, R. et al., Investigation of proton–proton short-range correlations via the $^{12}\text{C}(e,e'pp)$ reaction. *Phys. Rev. Lett.* **99**, 072501 (2007).

Resonant domain–wall–enhanced tunable microwave ferroelectrics

Zongquan Gu^{1,2}, Shishir Pandya³, Atanu Samanta⁴, Shi Liu⁵, Geoffrey Xiao¹, Cedric J. G. Meyers⁶, Anoop R. Damodaran³, Haim Barak⁴, Arvind Dasgupta³, Sahar Saremi³, Alessia Polemi¹, Liyan Wu⁷, Adrian A. Podpirka¹, Alexandria Will–Cole¹, Christopher J. Hawley¹, Peter K. Davies⁷, Robert A. York⁶, Ilya Grinberg⁴, Lane W. Martin^{3,8} & Jonathan E. Spanier^{1,2,9*}

Ordering of ferroelectric polarization¹ and its trajectory in response to an electric field² are essential for the operation of non-volatile memories³, transducers⁴ and electro-optic devices⁵. However, for voltage control of capacitance and frequency agility in telecommunication devices, domain walls have long been thought to be a hindrance because they lead to high dielectric loss and hysteresis in the device response to an applied electric field⁶. To avoid these effects, tunable dielectrics are often operated under piezoelectric resonance conditions, relying on operation well above the ferroelectric Curie temperature⁷, where tunability is compromised. Therefore, there is an unavoidable trade-off between the requirements of high tunability and low loss in tunable dielectric devices, which leads to severe limitations on their figure of merit. Here we show that domain structure can in fact be exploited to obtain ultralow loss and exceptional frequency selectivity without piezoelectric resonance. We use intrinsically tunable materials with properties that are defined not only by their chemical composition, but also by the proximity and accessibility of thermodynamically predicted strain-induced, ferroelectric domain-wall variants⁸. The resulting gigahertz microwave tunability and dielectric loss are better than those of the best film devices by one to two orders of magnitude and comparable to those of bulk single crystals. The measured quality factors exceed the theoretically predicted zero-field intrinsic limit owing to domain-wall fluctuations, rather than field-induced piezoelectric oscillations, which are usually associated with resonance. Resonant frequency tuning across the entire L, S and C microwave bands (1–8 gigahertz) is achieved in an individual device—a range about 100 times larger than that of the best intrinsically tunable material. These results point to a rich phase space of possible nanometre-scale domain structures that can be used to surmount current limitations, and demonstrate a promising strategy for obtaining ultrahigh frequency agility and low-loss microwave devices.

Current telecommunication devices rely on our ability to tune the device frequency in the radiofrequency spectrum. The development of bulk- and thin-film-based acoustic wave filters, resonators and other devices over the last few decades has allowed cell phone miniaturization, antenna tuning and the development of current mobile telecommunication technology. Further advancement (for example, 5G and IoT technologies) requires an even more efficient use of the spectrum, necessitating the development of thin-film dielectrics with higher dielectric tunability, n , quality factor, Q , and figures of merit, motivating intense research and development efforts. In particular, extrinsic effects, such as defects, strain, interface and polar ordering, have been intensely investigated and recent advances have enhanced our understanding of how functional properties can be tailored, can evolve from symmetry breaking or can even be induced artificially^{9–12}. Extrinsic

enhancement of susceptibility from ferroelectric domain walls¹³ can be attained by strain engineering through the creation of domain-wall-rich films, whose extrinsic character allows dielectric properties not bounded by the intrinsic limits of the defect-free bulk. Nevertheless, for tunable dielectrics, polar domains have not been considered helpful and are generally equivalent to other crystal imperfections (for example, oxygen vacancies) that must be suppressed to achieve greater material quality and thus lower dielectric loss and higher figure-of-merit values¹⁴. Therefore, domain engineering has not been investigated for tunable dielectrics.

In this work, we investigated the dielectric response of strained $\text{Ba}_x\text{Sr}_{1-x}\text{TiO}_3$ (BST) films in the vicinity of the Curie temperature, T_C . The nearly isotropic free-energy–polarization landscape (and lower barrier to polarization rotation) of these materials is expected to lead to a rich phase diagram and a large response to an applied electric field.

Thermodynamic Ginzburg–Landau–Devonshire (GLD) model calculations support the hypothesis that large in-plane permittivity values can be obtained via in-plane domains. Application of the phenomenological GLD model permits calculation of in-plane strain (u_s)–temperature (T)–polarization (P) phase diagrams and of the dielectric permittivity (Fig. 1; see Methods and Supplementary Information), with a number of additional domain variants (‘superdomain’ phases¹⁵) predicted for BST films (Fig. 1a–g; Methods and Supplementary Information). We focus on BST with $x = 0.8$, whose phase diagram exhibits a vertex, where a number of domain-wall-variant phases are predicted to intersect near room temperature (Fig. 1a; red circle). On the basis of the close proximity and high accessibility of the different variants, we refer to this region of the phase diagram as a manifold-domain-wall-variant material (MDVM). Zero- and finite-field phase-field model calculations for three selected strain states (denoted in Fig. 1a by the yellow dashed lines I, II and III) confirm the expected c^+/c^- structure for a compressively strained film (I) and an in-plane domain structure for a film under moderate tensile strain (III) (Fig. 1a). Application of a moderately large field (0.1 MV cm^{-1} along $[100]$) leaves the domain structure in I (Fig. 1b) and III (Fig. 1f) essentially unchanged (Fig. 1c, g). For case II, which corresponds to the MDVM material, the aa_1/aa_2 domain-wall-variant structure at zero field (Fig. 1d) is predicted, suggesting the coexistence of multiple domain-wall variants, which is consistent with the material’s location in the phase diagram. Despite the softer three-dimensional potential-energy landscape of II compared with those of I and III, the domain structure is not eliminated at moderate fields (Fig. 1e), consistent with reports on epitaxial films in which domain structures cannot be eliminated¹⁶ under an applied electric field.

We find that the dielectric permittivity values for the MDVM-engineered films exceed the composition-specific state of the art for dielectric thin films: theoretically predicted values for zero-field

¹Department of Materials Science and Engineering, Drexel University, Philadelphia, PA, USA. ²Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA.

³Department of Materials Science and Engineering, University of California at Berkeley, Berkeley, CA, USA. ⁴Department of Chemistry, Bar-Ilan University, Ramat-Gan, Israel. ⁵Carnegie Institution for Science, Washington, DC, USA. ⁶Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA, USA. ⁷Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, USA. ⁸Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁹Department of Physics, Drexel University, Philadelphia, PA, USA. *e-mail: spanier@drexel.edu

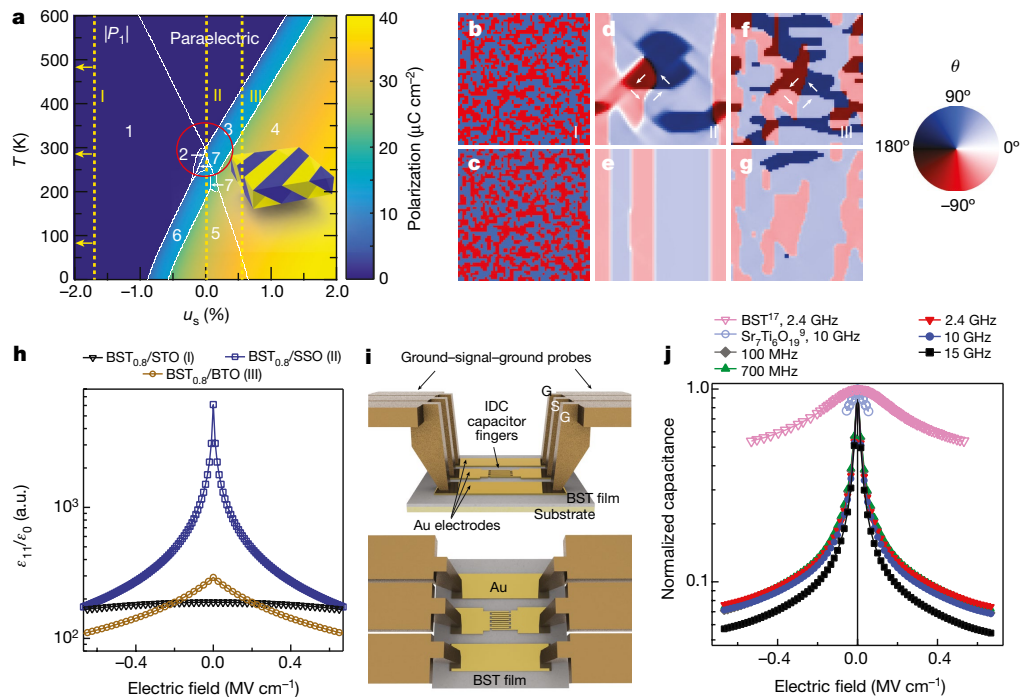


Fig. 1 | Design of a MDVM and its microwave dielectric tunability.

a, Thermodynamic landscape of in-plane polarization favouring domain-wall variants, showing the calculated average value of in-plane polarization, $|P_1|$, as a function of temperature, T , and in-plane strain, u_s , for $\text{Ba}_{0.8}\text{Sr}_{0.2}\text{TiO}_3$. Arabic numerals denote various thermodynamically predicted domain-wall-variant structures (see Supplementary Information). Dashed lines denote the strain states I, II and III, where state I lies outside the plotted range, as indicated by the yellow arrows. White lines denote the boundaries between domain-wall-variant phases. Shown in the inset is an illustration of an example domain-wall variant, where yellow and purple colours correspond to regions of the film with different polarization directions. **b–g**, Phase-field simulations of domain structure for equivalent strains and corresponding non-degenerate (I and III) and degenerate (II) domain-variant phase states in BaTiO_3 at about $10\text{ K} < T_C$ under zero field (**b**, **d**, **f**) reveal high out-of-plane (**b**) and in-plane (**d**, **f**) domain-wall densities that persist even under a moderate in-plane field of 0.1 MV cm^{-1} applied along $[100]$ (**c**, **e**, **g**). **b**, **c**, Maps of plane-normal

polarization P_3 , where blue and red colours correspond to $+30\text{ }\mu\text{C cm}^{-2}$ and $-30\text{ }\mu\text{C cm}^{-2}$, respectively. **d–g**, Maps of in-plane local polarization direction, as denoted by the white arrows. θ is the angle between $[100]$ and the sum of the two in-plane components, $P_1 + P_2$ (see Supplementary Information). **h**, The effect of proximity to this domain-phase-variant degeneracy point; displayed is the theoretically predicted in-plane quasi-static-field tunability of the relative dielectric permittivity, ϵ_{11}/ϵ_0 , in $\text{Ba}_{0.8}\text{Sr}_{0.2}\text{TiO}_3$ ($\text{BST}_{0.8}$) films on SrTiO_3 (STO), on SmScO_3 (SSO) and on BaTiO_3 (BTO), corresponding to strain states I, II and III, respectively, calculated using the GLD model. a.u., arbitrary units. **i**, Illustration of the experimental two-port co-planar geometry used to measure the microwave dielectric response of the BST epitaxial film (situated on the substrate), using ground (G) and source (S) input and output probes through the Au IDC finger electrodes. **j**, Measured in-plane field tuning of in-plane normalized capacitance at selected frequencies for a 400-nm-thick MDVM film sample ($\text{Ba}_{0.8}\text{Sr}_{0.2}\text{TiO}_3/\text{SmScO}_3(110)$), compared with those of epitaxial paraelectric BST^{17} and $\text{Sr}_7\text{Ti}_6\text{O}_{19}^9$ films.

relative dielectric permittivity ϵ_{11}/ϵ_0 easily exceed 10,000, reaching 10^5 for selected combinations. Higher permittivity promotes enhanced dielectric and capacitance tunability, n ($n = \epsilon_{r,\text{max}}/\epsilon_{r,\text{min}} = C_{\text{max}}/C_{\text{min}}$, where ϵ_r is the real part of the dielectric permittivity and C_{max} and C_{min} are the capacitances at zero field and with an applied electric field E , respectively), aided by proximity to the phase boundary. The theoretically calculated quasi-static in-plane tunability of MDVM films can be remarkably large. For example, an $x = 0.8$ film coherently strained on $\text{SmScO}_3(110)$ ($u_s \approx 0.05\%$, case II) is predicted to have tunability $n(E_1) > 20$ at $E_1 = 0.3\text{ MV cm}^{-1}$, whereas for films deposited on SrTiO_3 (I) and BaTiO_3 (III) n is considerably lower (Fig. 1h).

Experimental results support the GLD theory predictions. Epitaxial $x = 0.8$ films, 100 nm and 400 nm thick, were deposited on $\text{SmScO}_3(110)$ by pulsed-laser deposition and were characterized using a variety of techniques (Methods, Supplementary Information). Compared with the bulk, the smaller out-of-plane lattice parameters in our films favour in-plane domain formation, and plane-normal and lateral dual-amplitude resonance tracking (DART) piezoresponse force microscopy confirms the presence of in-plane oriented domains, with domain walls aligned along $[100]$ or $[010]$, consistent with the $aa_1/aa_2/aa_1/aa_2$ domain structure (Supplementary Information). Voltage-dependent capacitance data in the co-planar geometry (Fig. 1i, Methods) at selected frequencies across the measurement range demonstrate high tunability at modest fields (Fig. 1j), in agreement with our calculations, which persists beyond 20 GHz. This tunability, even

at equivalent fields, is considerably greater than the current state of the art in films grown by molecular beam epitaxy, including Ruddlesden–Popper $\text{Sr}_7\text{Ti}_6\text{O}_{19}^9$ and $(\text{Ba},\text{Sr})\text{TiO}_3^{17}$ (Fig. 1j). Remarkably, $n(f)$ remains greater than 13 (at 0.67 MV cm^{-1}) almost throughout the entire frequency range studied, peaking at $n \approx 18.5$ at 15.2 GHz (Supplementary Fig. 10). The deposited films also exhibit low losses (high Q values), in contrast to the high losses that usually accompany high tunabilities. MDVM films exhibit low Q at zero field, but large Q ($Q(f) \approx 1,200$, frequency-averaged in 0.1–20 GHz) at maximum field. At the highest applied field, Q ranges generally between 10^2 and 10^3 over the range 2–10 GHz (Supplementary Fig. 11).

A closer examination reveals extraordinary features in thinner films: field–frequency combinations for which Q oscillates with frequency easily exceed the frequency-dependent bulk intrinsic limit for BaTiO_3 near T_C (less than about 10^3 ; Supplementary Fig. 17), reaching, and even exceeding, 10^5 (Fig. 2; Supplementary Information). These Q values are much greater than the best ones reported for intrinsically tunable film materials^{17,18} (including ferroelectrics considered for high- Q dielectrics^{18–20}) and greater than those of AlN films^{21–24}, which are the leading non-ferroelectric (that is, not intrinsically tunable) piezoelectrics. They are comparable, in fact, to values measured for bulk single-crystal quartz²⁵, sapphire²⁶ and ZnO²⁷. The field dependence of the resonance frequency, $f_r(E)$, shows exceptional variation across one frequency decade, spanning the L (1–2 GHz), S (2–4 GHz) and C (4–8 GHz) bands and extending into the X band (8–12 GHz), all in a

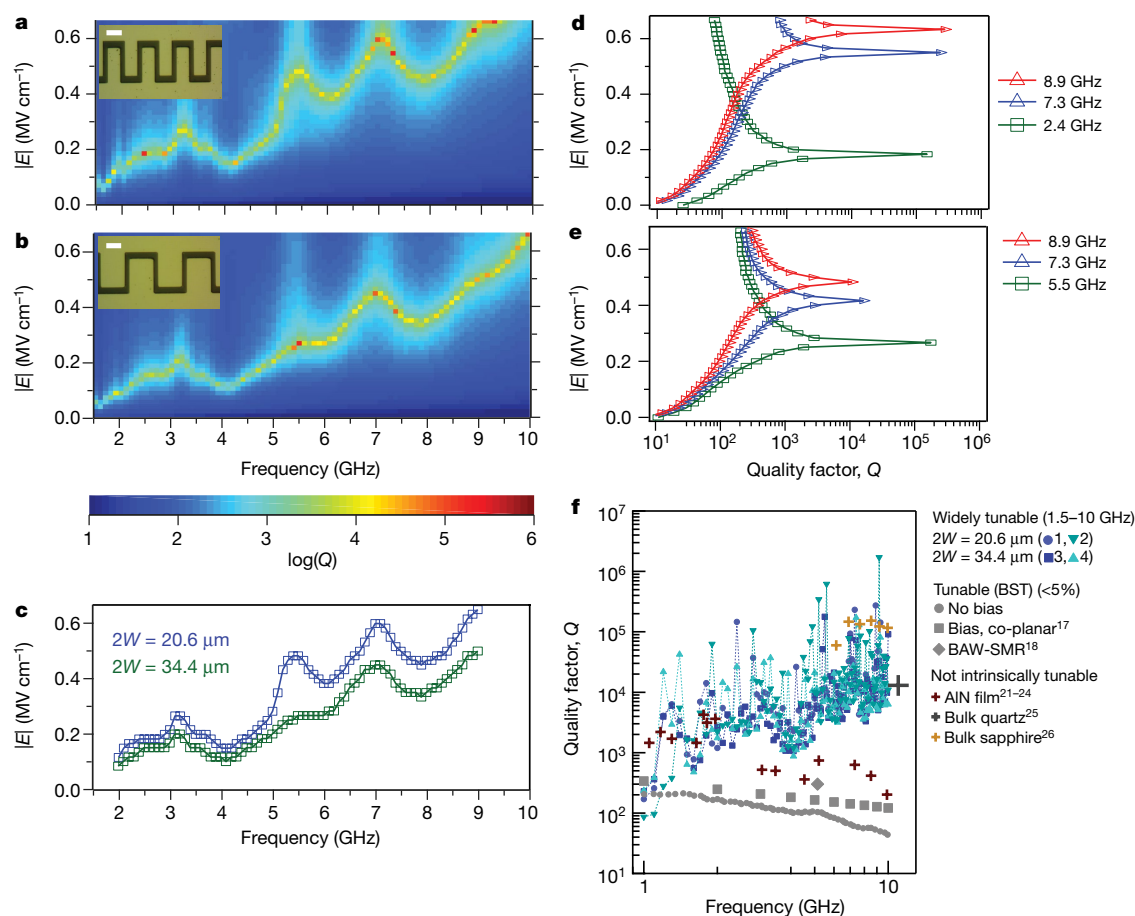


Fig. 2 | Microwave spectroscopy results revealing field-dependent resonant domain-wall spectral signatures of ultralow loss and tunable resonant performance. **a, b**, Experimentally determined Q as a function of frequency and d.c. in-plane bias field for a 100-nm-thick film with ten (**a**) and six (**b**) IDC electrode pairs. The period $2W$ is defined by the electrode finger width and the inter-electrode spacing. The insets show optical micrographs of the IDC electrode pairs (scale bar, 10 μm). **c**, Frequency and electric field values for which $Q(f)$ peaks are obtained for the devices in **a** (blue; $2W = 20.6 \mu\text{m}$) and **b** (green; $2W = 34.4 \mu\text{m}$). The plot shows that the spectra of the voltage-dependent frequencies for which the resonant $Q(f)$ peaks occur are essentially the same and that the resonant frequency can be bias-tuned by about 400%, from around 2 GHz at about 0.1 MV cm^{-1} to 10 GHz at about 0.67 MV cm^{-1} ,

with $10^3 \lesssim Q \lesssim 10^6$. **d, e**, Representative traces of d.c.-field-dependent Q at different frequencies for the devices shown in **a** and **b**, respectively. **f**, Peak Q values were collected at 100 frequencies in four distinct devices: the devices shown in **a** (1; blue circles) and **b** (3; blue squares) and two additional devices (2, 4) with the same characteristics as devices 1 and 3. The data show an increase of more than one order of magnitude over approximately one frequency decade, deviating strongly from the usual f^{-1} scaling law. Shown for comparison are the highest values reported for piezoelectric resonators made of bulk single-crystal quartz²⁵, sapphire²⁶ and AlN film^{21–24}, none of which is intrinsically tunable (each point represents an individual device), and values reported for intrinsically tunable BST films^{17,18}, including a bulk acoustic-wave solidly mounted resonator (BAW-SMR) film.

single device. The commutation quality factor²⁸, $\text{CQF}(f) = [n(f) - 1]^2 Q(0, f)Q(E, f)/n(f)$, a key metric that incorporates $n(E)$ and $Q(E)$, shows values that are greater than those of the best reported BST films¹⁷ (Supplementary Information).

Bulk dielectric and film resonators rely on electromechanical coupling of microwave power through piezoelectric oscillations, which appear as resonant and anti-resonant features that can be voltage-tuned by less than 4.5% in the best tunable materials²⁹. Considering the change in the piezoelectric coupling coefficient, the calculated bias-field dependence of the resonance and anti-resonance frequencies of in-plane piezoelectric oscillations for $\text{Ba}_{0.8}\text{Sr}_{0.2}\text{TiO}_3$ in our experimental geometry amounts to not more than about 3% for $0\text{--}0.6 \text{ MV cm}^{-1}$ (Supplementary Information)—hundreds of times lower than that observed in our devices. Furthermore, the design of piezoelectric resonators using in-plane piezoelectric oscillations operating at fundamental (or higher-mode) frequency relies on interdigitated capacitor (IDC) electrode periodicity^{23,30}. Although we cannot rule out potential contributions from other possible mechanisms (for example, piezoelectric oscillations), devices that differ in electrode finger width yield spectra that are essentially the same (Fig. 2c), further demonstrating that

it is highly unlikely that piezoelectric oscillations cause the observed spectrum.

We now investigate the origin of the unusual experimentally observed Q spikes using the data obtained from molecular dynamics simulations of a model BaTiO_3 (BTO) system (Methods). The analytical theory of the intrinsic dielectric response of a ferroelectric material¹⁴ predicts a $1/f$ dependence for $Q(f)$, as is also found in our single-domain molecular dynamics simulations (Supplementary Information), indicating that the unusual f -dependence of Q is due to extrinsic effects. Examination of static dielectric response shows that the peak dielectric constant value is observed in the ferroelectric phase (Supplementary Information). Such a domain-wall-driven shift of the dielectric response peak to the ferroelectric phase was previously observed experimentally in BaTiO_3 in the ferroelectric phase close to T_C ^{13,31}. In addition, reversible domain-wall oscillations are also found to lead to the peak dielectric constant in the ferroelectric phase for the model aa_1/aa_2 domain-wall supercell in our molecular dynamics simulations (Fig. 3a, Supplementary Information). The main reason for this domain-wall contribution to the dielectric permittivity is the existence of very low-energy modes localized on the two-dimensional

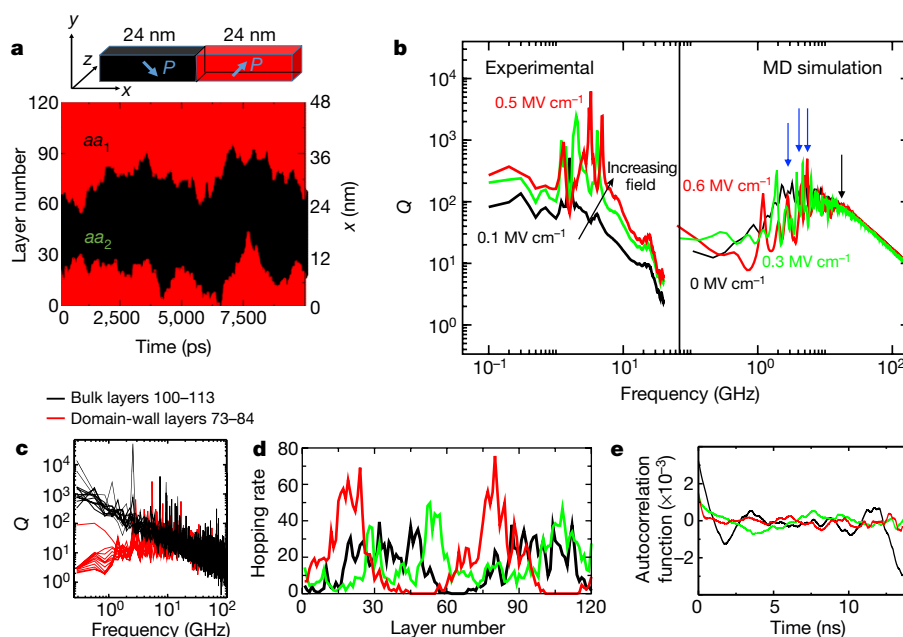


Fig. 3 | Molecular dynamics simulation of Q . **a**, Illustration of the molecular dynamics supercell and domain fluctuations at $E_x = 0.6 \text{ MV cm}^{-1}$, with the $P_y > 0$ domain shown in black (domain type aa_2) and the $P_y < 0$ domain shown in red (domain type aa_1). **b**, Experimental Q (left) and Q_y obtained by molecular dynamics (MD) simulations (right) for the aa_1/aa_2 domain structure. Experimental data are shown for $E = 0.09 \text{ MV cm}^{-1}$ (black), $E = 0.25 \text{ MV cm}^{-1}$ (green) and $E = 0.5 \text{ MV cm}^{-1}$ (red). Molecular dynamics simulation data are shown for $E = 0 \text{ MV cm}^{-1}$ (black), $E = 0.3 \text{ MV cm}^{-1}$ (green) and

$E = 0.6 \text{ MV cm}^{-1}$ (red). **c**, $Q(f)$ for the bulk-like layers 100–113 (black) and domain-wall layers 73–84 (red), obtained from molecular dynamics simulations at $E = 0.6 \text{ MV cm}^{-1}$. **d**, Hopping rates for individual layers of the $120 \times 10 \times 10$ supercell for $E = 0 \text{ MV cm}^{-1}$ (black), $E = 0.3 \text{ MV cm}^{-1}$ (green) and $E = 0.6 \text{ MV cm}^{-1}$ (red). **e**, Autocorrelation functions for the total polarization time, obtained from molecular dynamics simulations for $E = 0 \text{ MV cm}^{-1}$ (black), $E = 0.3 \text{ MV cm}^{-1}$ (green) and $E = 0.6 \text{ MV cm}^{-1}$ (red).

domain walls. The strong impact of domain-wall oscillations on the dielectric response and the presence of a high density of domain walls in our sample superdomain state suggest that these oscillations may also be the cause of $Q(f)$ oscillations.

To understand the relationships between the reversible domain-wall dynamics and $Q(f)$, we perform long (14 ns) simulations using a model system containing two aa_1/aa_2 domain walls in a $120 \times 10 \times 10$ supercell (Fig. 3a) at 50 K below the ferroelectric–paraelectric transition temperature and then obtain $Q(f)$ from the fluctuations of the total polarization of the supercell. We choose this size because at this domain length (24 nm) a clear distinction is observed between the domain wall and the bulk-like regions in the sample, as can be seen in Fig. 3a. Additionally, GLD theory predicts that the domain size should be of the order of 30 nm (Supplementary Information).

Comparison of the experimental and molecular-dynamics-obtained $Q(f)$ shows several similar features (Fig. 3b). First, at zero direct current (d.c.) bias, the linear or almost linear rise in Q with decreasing f is succeeded by flattening of $Q(f)$ with gentle oscillations, owing to the onset of relaxation at about 18 GHz (black arrow). This observation is in agreement with the expectation that the presence of domain walls leads to higher loss and lower Q , as can be seen from the much lower Q in the low- f region (less than 2 GHz for the experiment and less than 18 GHz for the molecular dynamics simulation) than that expected from the intrinsic $1/f$ dependence of Q . Second, at higher bias, peaks above the baseline appear at certain frequencies (blue arrows), with the $Q(f)$ curve shifting to higher frequencies with higher d.c. bias. Finally, a greater number of narrow peaks are observed in $Q(f)$ at higher bias. The $Q_y(f)$ results of the molecular dynamics simulations for $E = 0 \text{ MV cm}^{-1}$ and $E = 0.6 \text{ MV cm}^{-1}$ are qualitatively similar to the experimental $Q(f)$ data for $E = 0.09 \text{ MV cm}^{-1}$ and $E = 0.25 \text{ MV cm}^{-1}$, respectively, albeit at higher frequencies owing to the difference between the experimental BST and the computational BTO systems (Fig. 3b). The uniform shift to higher Q values with higher d.c. bias is not observed for molecular dynamics simulations, and this difference is probably due to the

difference between the simple model used in the molecular dynamics simulations and the much more complex E -field profile in the experimental samples.

Analysis of the $Q(f)$ curves of individual layers shows that the bulk-like layers (that is, layers in the middle of the domain that do not show switching) exhibit bulk-like $1/f$ dependence of Q , whereas the domain-wall layers exhibit $Q(f)$ spikes and a flattening out of the $Q(f)$ at low f , similar to the experimentally observed data and the $Q(f)$ obtained computationally for the total system (Fig. 3c). Comparison between the autocorrelation functions of the bulk-like and domain-wall layers (Supplementary Information) shows that the former exhibits the normal behaviour of rapid decay followed by small fluctuations around 0, whereas the latter shows a slow decay and large oscillation amplitude and period. This is due to the much larger fluctuations of the polarization P of the domain-wall layer between the two sides of the double-well potential compared to the oscillations of P inside a well. Therefore, domain-wall fluctuations dominate the dielectric response at low f .

Analysis of the polarization switching (from $-P_y$ to $+P_y$ and vice versa) rates for individual layers in the supercell shows that hopping rates increase with increasing d.c. bias (Fig. 3d), which can also be seen from the oscillations of the autocorrelation functions over the total polarization time (Fig. 3e). Therefore, the application of the d.c. bias accelerates the rate of domain-wall oscillations and leads to the shift of the $Q(f)$ curves to higher f . With no domain-wall oscillations, a bulk-like $1/f$ -dependence spectrum is obtained, whereas for slow domain-wall hopping a relaxation-driven flattening of $Q(f)$ is observed with gentle oscillations, and sharp peaks are obtained for faster hopping. This strongly suggests that the experimental $Q(f)$ spectrum with gentle oscillations at zero bias is due to the slow oscillations of the high domain-wall density, and the experimentally observed sharp peaks in $Q(f)$ are due to the acceleration of domain-wall hopping by the application of the d.c. bias.

To show that the domain-wall fluctuation mechanism alone can give rise to the observed sharp $Q(f)$ peaks, we perform stochastic

simulations using a simple model of coupled bistable oscillators with a domain wall (Supplementary Information). We find that domain-wall position oscillations and $Q(f)$ profiles qualitatively similar to those obtained in the molecular dynamics simulations can be obtained by adjusting the double-well parameters of the oscillators (Supplementary Information), demonstrating that domain-wall oscillations can give rise to the observed sharp variation in $Q(f)$. We note that bistable anharmonic oscillators are known to exhibit a variety of resonance behaviours, such as enhanced signal-to-noise ratio at certain frequencies, due to the fluctuation of the system between the two sides of the double well³². Thus, the behaviour of the MDVM BST film appears to be another example of this class of phenomena.

The hypothesis that the domain-wall position fluctuations give rise to the anomalous $Q(f)$ observed at high static bias explains why such characteristics have not been previously observed for $Q(f)$. To obtain $Q(f)$ oscillations, a large domain-wall density corresponding to domain size of less than 100 nm is necessary because otherwise the high Q arising from the domain walls will be averaged out by the normal behaviour of the bulk of the domain. In addition, this effect is likely to appear only close to T_C , where the thickness of the domain wall is larger and the barrier to switching is very low, enabling hopping of the domain-wall layer between the two P_y orientations at gigahertz frequencies. At lower T , the energy barrier for switching the P_y value of the layer is too high, so that the time necessary to cross the barrier between the two P_y states is too long; therefore, high Q would be observed only at frequencies in the megahertz range or below, where such an effect may not be apparent owing to the high Q of the bulk dielectric response at such low frequencies. Finally, films of very high quality are necessary to observe these effects, because variation in the frequencies of the very-low-dielectric-loss resonance due to defects, grain boundaries and compositional variations would lead to averaging out of the low loss and disappearance of the high- Q peaks.

The product Qf is one of the most frequently cited metrics for all dielectric microwave resonators, where acoustic attenuation parameterized by $\alpha \propto f^2$ in the Akhiezer limit³³ for phonon–phonon scattering leads to Qf equaling a material-specific constant. We note that in our experimental films, Qf deviates from the usual monotonic $Q(f)$ dependence for $1 \text{ GHz} < f_r < 10 \text{ GHz}$, showing a strong increase in this range. This suggests that the effective scattering rate due to thermal phonons is much lower than f_r , providing additional experimental evidence that our domain-wall-enhanced resonant films overcome intrinsic losses in this range. Meanwhile, our simulations of BTO indicate that the expected frequency band of voltage-tuned domain-wall resonances is material-specific and can be higher than that experimentally observed for BST. Thus, our results show that extrinsically driven MDVM tunable dielectric materials exhibit a quality factor that exceeds the intrinsic limit near T_C without piezoelectric oscillations, and are promising for achieving similar values of Q at a wider range of frequencies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0434-2>.

Received: 8 August 2017; Accepted: 26 June 2018;

Published online 20 August 2018.

- Merz, W. J. Domain formation and domain wall motions in ferroelectric BaTiO₃ single crystals. *Phys. Rev.* **95**, 690–698 (1954).
- Xu, R., Liu, S., Grinberg, I., Karthik, J., Damodaran, A. R., Rappe, A. M. & Martin, L. W. Ferroelectric polarization reversal via successive ferroelastic transitions. *Nat. Mater.* **14**, 79–86 (2015).
- Chanthbouala, A. et al. Solid-state memories based on ferroelectric tunnel junctions. *Nat. Nanotechnol.* **7**, 101–104 (2012).
- Murali, P. Ferroelectric thin films for micro-sensors and actuators: a review. *J. Micromech. Microeng.* **10**, 136–146 (2000).
- Wessels, B. W. Ferroelectric epitaxial thin films for integrated optics. *Annu. Rev. Mater. Res.* **37**, 659–679 (2007).
- Arlt, G., Böttger, U. & Witte, S. Dielectric dispersion of ferroelectric ceramics and single crystals at microwave frequencies. *Ann. Phys.* **506**, 578–588 (1994).

- York, B. in *Multifunctional Adaptive Microwave Circuits and Systems* (eds Steer, M. & Palmer, W. D.) Ch. 6 (SciTech Publishing, Raleigh, 2006).
- Pertsev, N. A., Zembilgotov, A. G. & Tagantsev, A. K. Effect of mechanical boundary conditions on phase diagrams of epitaxial ferroelectric thin films. *Phys. Rev. Lett.* **80**, 1988–1991 (1998).
- Lee, C.-H. et al. Exploiting dimensionality and defect mitigation to create tunable microwave dielectrics. *Nature* **502**, 532–536 (2013).
- Damodaran, A. R., Breckenfeld, E., Chen, Z., Lee, S. & Martin, L. W. Enhancement of ferroelectric Curie temperature in BaTiO₃ films via strain-induced defect dipole alignment. *Adv. Mater.* **26**, 6341–6347 (2014).
- Damodaran, A. R. et al. Nanoscale structure and mechanism for enhanced electromechanical response of highly strained BiFeO₃ thin films. *Adv. Mater.* **23**, 3170–3175 (2011).
- Prosandeev, S., Yang, Y., Paillard, C. & Bellaiche, L. Displacement current in domain walls of bismuth ferrite. *npj Comput. Mater.* **4**, 8 (2018).
- Wang, Y. L., Tagantsev, A. K., Damjanovic, D. & Setter, N. Giant domain wall contribution to the dielectric susceptibility in BaTiO₃. *Appl. Phys. Lett.* **91**, 062905 (2007).
- Tagantsev, A. K., Sherman, V. O., Astafiev, K. F., Venkatesh, J. & Setter, N. Ferroelectric materials for microwave tunable applications. *J. Electroceram.* **11**, 5–66 (2003).
- Matzen, S. et al. Super switching and control of in-plane ferroelectric nanodomains in strained thin films. *Nat. Commun.* **5**, 4415 (2014).
- Griggio, F. et al. Composition dependence of local piezoelectric nonlinearity in (0.3)Pb(Ni_{0.33}Nb_{0.67})O₃–(0.7)Pb(Zr_{1–x}Ti_x)O₃ films. *J. Appl. Phys.* **110**, 044109 (2011).
- Meyers, C. J. G., Freeze, C. R., Stemmer, S. & York, R. A. (Ba, Sr)TiO₃ tunable capacitors with RF commutation quality factors exceeding 6000. *Appl. Phys. Lett.* **109**, 112902 (2016).
- Vorobiev, A., Gevorgian, S., Löffler, M. & Olsson, E. Correlations between microstructure and Q -factor of tunable thin film bulk acoustic wave resonators. *J. Appl. Phys.* **110**, 054102 (2011).
- Budimir, M., Damjanovic, D. & Setter, N. Extension of the dielectric tunability range in ferroelectric materials by electric bias field antiparallel to polarization. *Appl. Phys. Lett.* **88**, 082903 (2006).
- Rojac, T., Bencan, A., Drazic, G., Kosec, M. & Damjanovic, D. Piezoelectric nonlinearity and frequency dispersion of the direct piezoelectric response of BiFeO₃ ceramics. *J. Appl. Phys.* **112**, 064114 (2012).
- Zuo, C., Der Spiegel, J. V. & Piazza, G. 1.05-GHz cmos oscillator based on lateral-field-excited piezoelectric AlN contour-mode MEMS resonators. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **57**, 82–87 (2010).
- Gong, S., Kuo, N. K. & Piazza, G. GHz high- Q lateral overmoded bulk acoustic-wave resonators using epitaxial SiC thin film. *J. Microelectromech. Syst.* **21**, 253–255 (2012).
- Rinaldi, M., Zuniga, C., Zuo, C. & Piazza, G. Super-high-frequency two-port AlN contour-mode resonators for RF applications. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **57**, 38–45 (2010).
- Rinaldi, M., Zuniga, C. & Piazza, G. 5–10 GHz AlN contour-mode nanoelectromechanical resonators. In *2009 IEEE 22nd International Conference on Micro Electro Mechanical Systems* 916–919 (IEEE, 2009).
- Krupka, J., Tobar, M. E., Hartnett, J. G., Cros, D. & Le Floch, J. M. Extremely high- Q factor dielectric resonators for millimeter-wave applications. *IEEE Trans. Microw. Theory Tech.* **53**, 702–712 (2005).
- Hartnett, J. G., Tobar, M. E., Ivanov, E. N. & Krupka, J. Room temperature measurement of the anisotropic loss tangent of sapphire using the whispering gallery mode technique. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **53**, 34–38 (2006).
- Magnusson, E. B. et al. Surface acoustic wave devices on bulk ZnO crystals at low temperature. *Appl. Phys. Lett.* **106**, 063509 (2015).
- Vendik, I. B., Vendik, O. G. & Kollberg, E. L. Commutation quality factor of two-state switchable devices. *IEEE Trans. Microw. Theory Tech.* **48**, 802–808 (2000).
- Berge, J. & Gevorgian, S. Tunable bulk acoustic wave resonators based on Ba_{0.25}Sr_{0.75}TiO₃ thin films and a HfO₂/SiO₂ Bragg reflector. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **58**, 2768–2771 (2011).
- Gevorgian, S. S., Tagantsev, A. K. & Vorobiev, A. K. *Tunable Film Bulk Acoustic Wave Resonators* (Springer, New York, 2013).
- Hoshina, T. et al. Domain size effect on dielectric properties of barium titanate ceramics. *Jpn. J. Appl. Phys.* **47**, 7607–7611 (2008).
- Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–287 (1998).
- Akhiezer, A. On the absorption of sound in solids. *J. Phys. USSR* **1**, 277 (1939).

Acknowledgements Work at Drexel University and the University of California at Berkeley was supported in part by the US National Science Foundation (NSF) and the Semiconductor Research Corporation under the ‘Nanoelectronics in 2020 and Beyond’ programme grant number DMR 1124696 and by the Materials Science Division of the US Army Research Office (ARO). Z.G. and G.X. acknowledge support from the ARO under grant number W911NF-14-1-0500. A.P. and A.A.P. acknowledge support from the NSF under grant number IIP 1549668. A.W.-C. acknowledges support from the NSF under grant number DMR 1608887. C.J.H. acknowledges support from the Office of Naval Research under grant number N00014-15-1-2170. J.E.S. acknowledges support from the Air Force Office of Scientific Research under grant number FA9550-13-1-012. I.G., A.S., H.B., J.E.S. and G.X. acknowledge support from the NSF–BSF (US–Israel Binational Science

Foundation) joint programme under grant numbers BSF 2016637 and CBET 1705440. S.P. and A.R.D. acknowledge support from the ARO under grant number W911NF-14-1-0104. A.R.D. also acknowledges the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under award number DE-SC-0012375 for the development of the BST materials. A.D. acknowledges support from the NSF under grant number DMR 1708615. S.S. acknowledges support from the NSF under grant number DMR 1608938. L.W.M. acknowledges support from the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract number DE-AC02-05-CH11231: Materials Project programme KC23MP for the development of new functional materials. R.A.Y. and C.J.G.M. acknowledge support from ARO under grant number W911NF-14-1-0335. Numerical GLD and phase-field simulations were carried out on Proteus, a computer cluster supported by the Drexel University Research Computing Facility.

Reviewer information *Nature* thanks S. Prosandeev, A. Vorobiev and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.G. and J.E.S. conceived the idea and, together with I.G., proposed the mechanism for high Q . Z.G. and J.E.S. developed and implemented the GLD model; G.X., Z.G. and J.E.S. implemented the phase-field model; and A.S., S.L. and I.G. carried out the molecular dynamics simulations and analysis of the molecular dynamics data. H.B. and I.G. developed the stochastic oscillator model. S.P., A.R.D., A.D., Z.G., L.W.M. and J.E.S. designed the

growth experiments, and Z.G., I.G. and J.E.S. formulated the interpretation of the experimental data. A.D., A.R.D., S.P. and Z.G. carried out the film synthesis and its optimization. L.W. and P.K.D. produced the solid-state sources. S.P., A.D. and Z.G. carried out X-ray diffraction measurements, X-ray reflectivity analysis and reciprocal space mapping. Z.G., A.D. and S.P. carried out piezoresponse force microscopy analysis. C.J.G.M. designed the microwave devices, carried out film processing and device fabrication and performed microwave measurements and analysis, which were supervised by R.A.Y. Assistance in microwave device design, fabrication, measurements and analysis was provided by A.P., A.A.P. and A.W.-C. The theoretical, experimental and computational modeling aspects of the project were overseen together by I.G., R.A.Y., L.W.M. and J.E.S. All authors contributed to the data analysis. Z.G., I.G. and J.E.S. wrote the manuscript. Z.G., G.X., S.P., A.S., A.D., A.R.D., S.S., C.J.G.M., R.A.Y., I.G., L.W.M. and J.E.S. edited and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0434-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.E.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

GLD thermodynamic phase and field-dependent susceptibility calculations.

Accounting for the self-deformation of BTO and STO to accommodate the BST solid solution³⁴, the Gibbs free energy density, f_G , includes the Landau energy, f_L , the elastic energy, f_{ela} , and the electrostatic energy, f_{elec} . The total Helmholtz energy, f_{tot} , is obtained through the Lagrange transformation, and the clamped-thin-film boundary condition is applied to eliminate the strain and stress fields. By minimizing f_{tot} with respect to the order parameter polarization, P , the phase diagram is constructed as a function of the temperature, T , and substrate strain, u_s , with the paraelectric, tetragonal, orthorhombic and rhombohedral phases inherited from parent BTO in both single-domain and domain-wall-variant cases. The detailed derivation can be found in Supplementary Information.

Phase field calculations of the polydomains. The domain evolution under external d.c. bias is simulated using a phase-field model. The total free energy density f_{tot} includes the Landau (f_L), electrostriction (f_{es}), elastic (f_{ela}), electrostatic (f_{elec}) and gradient (f_{grad}) contributions. The time-dependent Ginzburg–Landau equation, $dP/dt = -L \times \delta \int_{\text{tot}} dV / \delta P$, where L is the kinetic coefficient, is solved in k space with periodic boundary conditions in the film plane and thin-film boundary conditions in the plane normal direction. Details can be found in Supplementary Information.

Pulsed laser deposition. BST targets were prepared by solid-state sintering of commercially available powders. $\text{SmScO}_3(110)$ substrates were attached to a heater by silver paste and placed several centimetres away from the BST target. Substrates were heated to 600 °C, and pulsed laser deposition was carried out using an excimer laser ($\lambda = 248$ nm) with a fluence of around 1.2 J cm^{-2} and a repetition rate of 3 Hz at an O_2 pressure of 20 mtorr. After reaching the expected thickness, the deposition was stopped, an O_2 flow of 1 atm was supplied into the chamber and the films were cooled to room temperature at a rate of 5°C min^{-1} .

Structural characterizations. X-ray diffraction and reciprocal space mapping about the pseudocubic (103) substrate plane were performed using a Panalytical MRD diffractometer. $\theta/2\theta$ scans (with minor umweg peaks) confirmed phase purity and indicated that each film was strained with the substrate.

Piezoresponse force microscopy. Normal and lateral piezoresponse force microscopy analysis was carried out at room temperature using the DARTTM mode as implemented on an Asylum Research MFP-3D scanning probe microscope. Each sample was scanned at 0° and 45° to eliminate the possibility of scan-angle-induced artefacts in the observed patterns.

Fabrication of test structures. Two-port IDC electrodes were fabricated on the samples by a lift-off metallization process using bi-layer photoresist. To begin, the samples were cleaned with ultrasonic agitation in acetone, followed by an isopropanol rinse. The resist stack consisted of $2.2 \mu\text{m}$ of polydimethylglutarimide (PMGI) polymer topped by $1.8 \mu\text{m}$ of negative photoresist (AZ nLOF-2020). The PMGI was baked for 5 min at 180°C and then cooled to 50°C at 5°C min^{-1} before the imaging resist was spun, then baked at 110°C and similarly cooled. The negative imaging resist was exposed through a photomask using projection lithography and developed in tetra-methyl ammonium hydroxide developer. After rinsing in deionized water, the sample was flood-exposed with deep ultraviolet light to scissor the cross-linked PMGI underlayer. The ultraviolet-scissored PMGI underlayer was developed using a tetra-ethyl ammonium hydroxide developer, in which the imaging resist is insoluble, yielding a retrograde profile in the resist sidewall. Finally, 15 nm of Ti was evaporated by electron-beam evaporation as an adhesion layer, followed by $1 \mu\text{m}$ of Au. The sample was then soaked in 1165 solvent overnight to complete the lift-off.

Microwave radiofrequency measurements. Devices were tested using a two-port microwave setup capable of applying large d.c. bias voltage, illustrated in Fig. 1i. Data were collected for the frequency and voltage dependence of the two-port complex reflection coefficients from 0.1 GHz to 40 GHz in 100-MHz steps and for an applied bias range from -200 V to $+200 \text{ V}$ (corresponding to $|E| \leq 0.67 \text{ MV cm}^{-1}$). Microwave radiofrequency measurements were performed using a four-port programmable vector network analyser (VNA; Agilent N5227A, 67 GHz) and a Cascade Microtech Summit 9K probe station equipped

with $150\text{-}\mu\text{m}$ - and $100\text{-}\mu\text{m}$ -pitch coplanar ground–signal–ground probes (Infinity I67-A-150-GSG-HC, Cascade Microtech). The system was cabled with flexible radiofrequency cables fitted with 1.85-mm V-connectors with a swept right angle at the probe end (4F0BX0BB0240, W.L. Gore and Associates). External high-voltage bias tees attached in-line at the probe heads allowed a d.c. bias voltage, supplied by an external Keithley 2634A source-measurement unit, to be applied to the IDC electrodes during the radiofrequency measurements of the frequency and voltage dependence of the two-port complex reflection coefficients. The measurement system was calibrated using a line–reflect–reflect–match algorithm. The calibration measurements were performed with WinCal XE software from Cascade Microtech using thru, short and $50\text{-}\Omega$ -load standards on an impedance standard substrate (101-109C, Cascade Microtech) and an open load standard formed by raising the probe station platen lever to lift the probes to more than $200 \mu\text{m}$ above the impedance standard substrate. Each frequency sweep was composed of three averaged sweeps with a system (intermediate frequency) bandwidth of 200 Hz, corresponding to a noise floor better than -100 dBm . The VNA power level was set to -15 dBm with a power slope of 0.01 dB GHz^{-1} to compensate for increased cable loss at higher frequencies. When collecting the data, voltage was stepped from zero to 5 V , -5 V , 10 V , -10 V , and so on, to $\pm 200 \text{ V}$. At each bias point, the voltage was held for 5 s before the VNA measurement was triggered. The total applied voltage was divided across the two device ports, referenced to port 1 so that for a given bias V , $V/2$ was applied to port 1 and $-V/2$ was applied to port 2, with polarities inverted for negative bias points (see Supplementary Information).

Molecular dynamics calculations. The simulations were performed in the constant-pressure constant-temperature ensemble using a Nose–Hoover thermostat and a Rahman–Parinello barostat with a 1 fs time step in a $10 \times 10 \times 10 \text{ BaTiO}_3$ supercell (5,000 atoms) for the single-domain calculations and in a $120 \times 10 \times 10 \text{ BaTiO}_3$ supercell (60,000 atoms) for domain simulations. A detailed description of the bond-valence potential is provided elsewhere³⁵. Briefly, this potential is parameterized to reproduce density functional theory energies and forces and reproduces the four phases (rhombohedral, orthorhombic, tetragonal and cubic) of BaTiO_3 . However, owing to the known underestimation of the ferroelectric–paraelectric T_C by density functional theory, the Curie temperature for the bulk bond-valence BaTiO_3 is 160 K , and it is 200 K for the strain conditions of the $120 \times 10 \times 10$ domain-wall sample simulations. Therefore, to study the dielectric response of the material in the paraelectric phase (and around T_C), we performed bond-valence molecular-dynamics simulations at 150 K , 160 K , 180 K and 200 K (10 K below T_C , at T_C , and at 20 K and 40 K above T_C , respectively). Once 10 ns of atomic trajectory data were obtained, we calculated the total dipole moment of the simulation supercell using the Born effective charges of the individual atoms and their instantaneous displacements from the high-symmetry positions. The total dipole moment was then used to calculate the autocorrelation function, which was then Fourier-transformed to obtain the real (ϵ_r) and imaginary (ϵ_{imag}) components of the dielectric constant and the quality factor.

Calculation of the hopping rates. To evaluate the average hopping rate of the average polarization of a layer in the xy plane in our supercell, we count the number of times, N , that the average polarization of the layer switched from a positive value at time t to a negative value at time $t + \delta t$, where $\delta t = 0.5 \text{ ps}$ is the interval between the printing of the atomic coordinates during the simulation. Because the total simulation time is 10 ns , the average rate of P switching for the layer is $N/10 \text{ ns}^{-1}$ or $N/10 \text{ GHz}$. This corresponds to the hopping rate of the layer between the two sides of the potential energy well.

Data availability. Data are available from the corresponding author upon reasonable request.

34. Shirokov, V. B., Yuzkuk, Yu. I., Dkhil, B. & Lemanov, V. V. Phenomenological theory of phase transitions in epitaxial $\text{Ba}_x\text{Sr}_{1-x}\text{TiO}_3$ thin films. *Phys. Rev. B* **79**, 144118 (2009).
35. Qi, Y., Liu, S., Grinberg, I. & Rappe, A. M. Atomistic description for temperature-driven phase transitions in BaTiO_3 . *Phys. Rev. B* **94**, 134308 (2016).

Sensitivity of atmospheric CO₂ growth rate to observed changes in terrestrial water storage

Vincent Humphrey^{1*}, Jakob Zscheischler¹, Philippe Ciais², Lukas Gudmundsson¹, Stephen Sitch³ & Sonia I. Seneviratne^{1*}

Land ecosystems absorb on average 30 per cent of anthropogenic carbon dioxide (CO₂) emissions, thereby slowing the increase of CO₂ concentration in the atmosphere¹. Year-to-year variations in the atmospheric CO₂ growth rate are mostly due to fluctuating carbon uptake by land ecosystems¹. The sensitivity of these fluctuations to changes in tropical temperature has been well documented^{2–6}, but identifying the role of global water availability has proved to be elusive. So far, the only usable proxies for water availability have been time-lagged precipitation anomalies and drought indices^{3–5}, owing to a lack of direct observations. Here, we use recent observations of terrestrial water storage changes derived from satellite gravimetry⁷ to investigate terrestrial water effects on carbon cycle variability at global to regional scales. We show that the CO₂ growth rate is strongly sensitive to observed changes in terrestrial water storage, drier years being associated with faster atmospheric CO₂ growth. We demonstrate that this global relationship is independent of known temperature effects and is underestimated in current carbon cycle models. Our results indicate that interannual fluctuations in terrestrial water storage strongly affect the terrestrial carbon sink and highlight the importance of the interactions between the water and carbon cycles.

Acquiring accurate estimates of the land carbon sink is a key requirement for monitoring global CO₂ emissions on a year-to-year basis⁸ and for reducing large uncertainties in projections of future carbon cycle–climate feedbacks^{9,10}. One critical aspect is to understand the sensitivity of the CO₂ growth rate (CGR) to natural climate variability. At the global scale, it was found that the interannual variability (IAV) of the CGR is coupled with the El Niño Southern Oscillation (ENSO) and more specifically with variations in mean tropical temperature^{3,4,6,11}. In addition, the role of water availability has been widely documented at the regional scale. Major droughts have been shown to cause drastic regional reductions in the land carbon sink^{12,13} and photosynthesis is limited by water scarcity over most of the globe¹⁴. Previous attempts to quantify the response of CGR IAV to water scarcity have used proxies to represent the amount of water available to ecosystems, such as yearly means of precipitation anomalies⁶, time-lagged and low-pass filtered monthly precipitation^{3,5} or standardized drought indices⁴. Although convenient to use, these proxies are limited because they consider only water inputs and either omit or model water losses due to evapotranspiration and runoff. From a process perspective, plants and microorganisms respond to the amount of water stored on land rather than to precipitation fluxes (Extended Data Fig. 1). We overcome these limitations by using direct satellite observations of terrestrial water storage (TWS) anomalies to investigate links between the carbon and water cycles.

From 2002 to 2017, the twin satellites of the Gravity Recovery and Climate Experiment (GRACE) have measured monthly anomalies of the Earth's gravity field⁷ that can be used to retrieve net changes in TWS, which include groundwater, soil moisture, surface waters, snow and water stored in the biosphere (see Methods). We isolate the monthly TWS IAV from GRACE by subtracting the mean seasonal cycle, and

remove the long-term trend using linear regression. Measurements of atmospheric CGR IAV from the National Oceanic and Atmospheric Administration (NOAA) are compared with the satellite-based TWS IAV over the overlapping period, revealing a significant negative correlation ($P < 0.05$ throughout, wherever significance is mentioned; the test is done with a moving block bootstrapping approach; see Methods) at both monthly ($r = -0.65$, $n = 158$) and yearly ($r = -0.85$, $n = 15$) scales (Fig. 1a, b). The sign of this relationship indicates that drier years, characterized by a negative anomaly in TWS, are associated with higher rates of atmospheric CO₂ growth and therefore a weakening of the land carbon sink (Fig. 1b). Composite mean TWS maps associated with high (Fig. 1c) and low (Fig. 1d) monthly CGR primarily reflect winter–spring water storage anomalies in South America and tropical regions in general. Given the relatively short observational record provided by the GRACE satellites, we investigate the robustness of this coupling and the associated spatial patterns using alternative estimates of TWS, which offer longer temporal coverage (Methods, Supplementary Tables 1 and 3). Although comparable, these estimates are based on model simulations that are considered less reliable than the actual GRACE observations. First, we use a statistical model of climate-driven water storage variability that is trained with GRACE observations (GRACE-REC)¹⁵ (Fig. 1a). We exclude the years following the eruption of Mt Pinatubo (1991–1993) in the Philippines^{16–18} (Methods), and find a significant negative coupling between GRACE-REC TWS and CGR at both monthly ($r = -0.59$, $n = 408$) and yearly ($r = -0.61$, $n = 34$) scales over the period 1980–2016 (Fig. 2a, b). Using TWS simulated by process-based land surface models, we tend to find lower correlations as we move from global hydrological models, which usually have the most complete or well calibrated representation of water reservoirs (WaterGAP¹⁹), to land surface models (GLDAS2-Noah²⁰) or dynamic global vegetation models (DGVMs), which often consider only root-zone soil moisture (TRENDY ensemble, version 3⁹). Nevertheless, these model estimates confirm the existence of a coupling between water storage and observed CGR. Unlike precipitation anomalies, water storage changes integrate the history of variations in both water supply and water demand over time. Therefore, looking at precipitation alone (with an optimal 4-month lag⁵) underestimates the strength of the coupling between water storage and carbon fluxes, in particular at the monthly scale (Fig. 2a, c). The strength of the link between CGR and water storage is comparable to that of the link between ENSO and CGR with a lag of about 4 months (ENSO leading CGR⁵). ENSO is a key mode of variability in global atmospheric circulation and is associated with large-scale fluctuations in precipitation patterns, which ultimately translate into water storage anomalies^{6,21} (Extended Data Fig. 3).

As documented in previous studies^{3,5}, the correlation between CGR and temperature is more pronounced in the tropical domain and on yearly timescales (Fig. 2c, d). Individual effects of temperature and water storage on CGR may be difficult to disentangle because these two drivers co-vary. Warmer years generally coincide with drier years (Fig. 3a), raising the question of whether the TWS signal might

¹Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland. ²Laboratoire des Sciences du Climat et de l'Environnement, CEA CNRS UVSQ, Gif-sur-Yvette, France.

³College of Life and Environmental Sciences, University of Exeter, Exeter, UK. *e-mail: vincent.humphrey@env.ethz.ch; sonia.seneviratne@ethz.ch

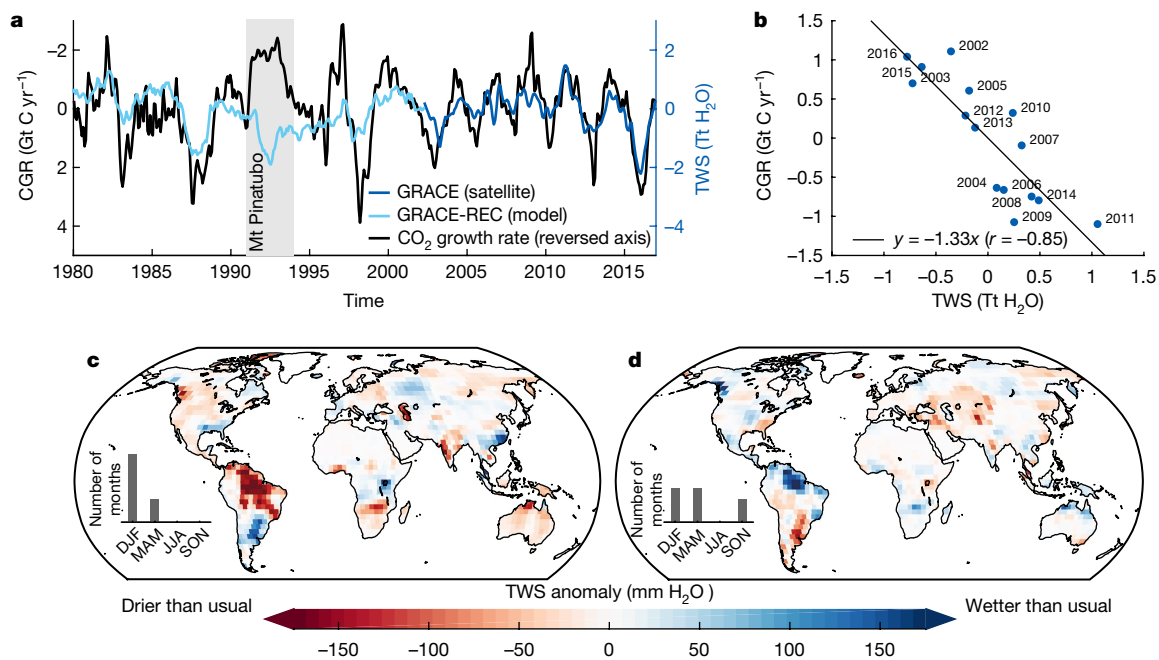


Fig. 1 | IAV in CGR and TWS. **a**, Monthly de-seasonalized and de-trended CGR, TWS from satellite observations (GRACE) and TWS from a statistical model (GRACE-REC¹⁵). The vertical axis is inverted for CGR so that positive (downwards) CGR anomalies indicate a weaker land carbon sink. A 6-month moving average was applied to GRACE data for

readability. **b**, Yearly CGR versus GRACE TWS anomalies. **c**, **d**, Composite mean TWS anomalies associated with the 5% highest (**c**) and 5% lowest (**d**) monthly CGR ($n = 8$; see Source Data for Fig. 1). Inset bar plots indicate the season of the corresponding months. Composites based on GRACE-REC show similar patterns (Extended Data Fig. 2).

implicitly contain some response to temperature. However, our results show that GRACE TWS can be almost entirely reconstructed from precipitation anomalies alone (Extended Data Fig. 4) with very little impact from temperature variability. Partial correlations indicate that the global CGR–TWS relationship remains significant after controlling for the effect of either global or tropical temperature (partial correlations r of -0.72 ; Fig. 3b, blue bars). This means that most of the information on CGR variations that is contained in TWS cannot be

found in temperature. In contrast, controlling for the effect of TWS strongly decreases partial correlations between CGR and temperature (Fig. 3b, orange bars). Using univariate linear regression (Methods), we find a global yearly sensitivity of -1.33 Gt (95% confidence interval (CI) spanning from -1.85 to -1.07 Gt) of carbon per year for each additional Tt of water stored on land (Fig. 3c). This corresponds to a ratio of roughly 1.3 g C yr⁻¹ per kg H₂O. When including both TWS and temperature in a bivariate regression, the sensitivity to TWS is

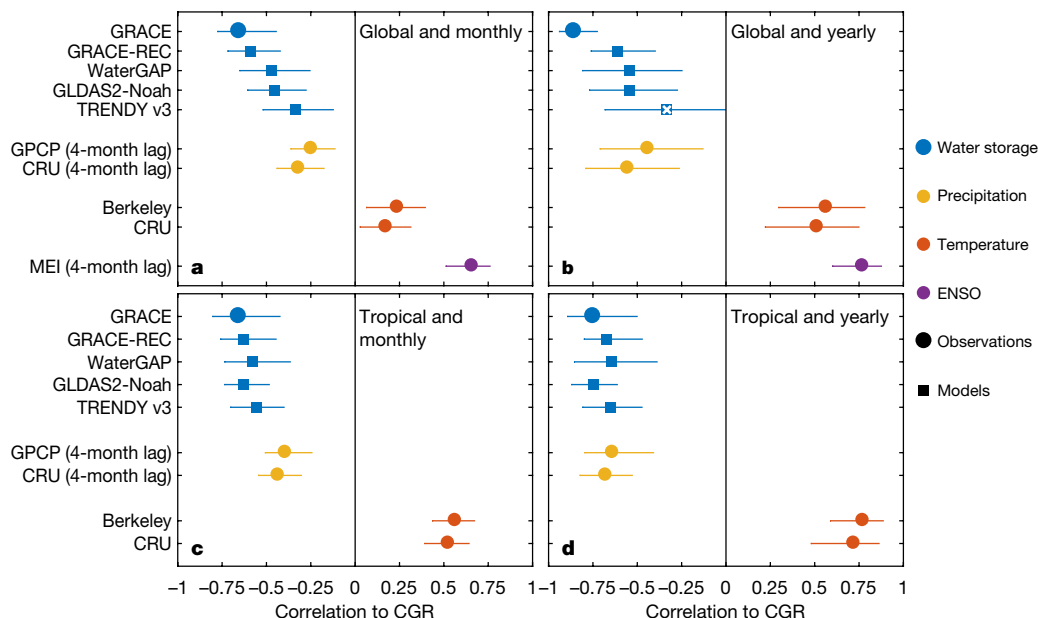


Fig. 2 | Correlations between CGR and meteorological drivers over different spatial domains at monthly and yearly scale. The years 1991–1993 affected by the eruption of Mt Pinatubo are excluded (Methods). Observations (circles) are distinguished from model-based estimates (squares). A white cross indicates a non-significant correlation

($P > 0.05$; Methods). Horizontal lines correspond to the 95% confidence interval of the correlation coefficient (Methods). The different products as well as the number of data points used to generate these results are listed in Supplementary Tables 1 and 3.

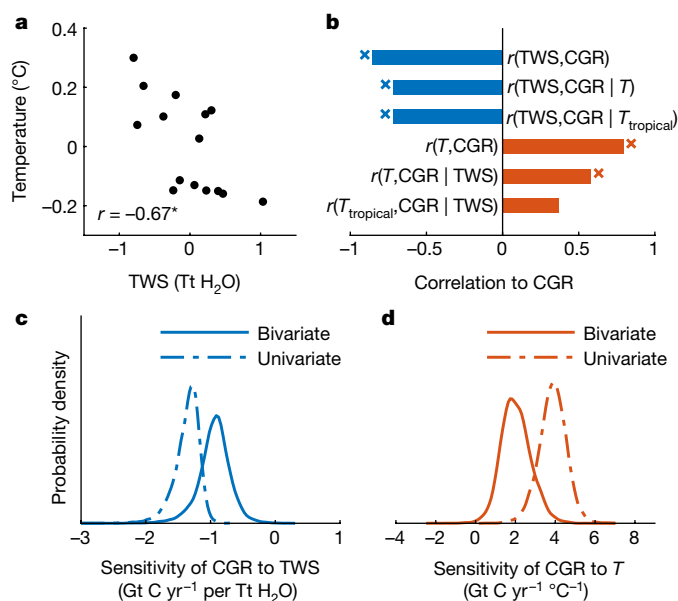


Fig. 3 | Confounding effects of water storage and temperature on correlations with CGR. **a**, Yearly co-variation between global mean GRACE TWS and global mean temperature over the period 2002–2016. **b**, Shown in blue, from top to bottom, is first the correlation between GRACE TWS and CGR, followed by partial correlations r between GRACE TWS and CGR after controlling for the effect of either global temperature T or tropical temperature T_{tropical} . Shown in orange is the correlation between CGR and T , and partial correlations between T (or T_{tropical}) and CGR after controlling for TWS. Significance ($P < 0.05$, $n = 15$; see Methods) is indicated with crosses. **c**, **d**, Probability distributions of the yearly sensitivities of CGR to TWS (**c**) and to T (**d**) in a univariate (dashed line) versus bivariate (solid line) regression (Methods).

reduced to $-0.93 \text{ Gt C yr}^{-1} \text{ per Tt H}_2\text{O}$ (CI -1.50 to $-0.48 \text{ Gt C yr}^{-1} \text{ per Tt H}_2\text{O}$, a 28% decrease). For temperature, the univariate sensitivity is $3.89 \text{ Gt C yr}^{-1} \text{ °C}^{-1}$ (CI 2.44 to $5.16 \text{ Gt C yr}^{-1} \text{ °C}^{-1}$) and is largely reduced in the bivariate case to $1.99 \text{ Gt C yr}^{-1} \text{ °C}^{-1}$ (CI 0.66 to $3.59 \text{ Gt C yr}^{-1} \text{ °C}^{-1}$, a 49% decrease), which is much lower than previous estimates^{2,4} (Fig. 3d).

Our findings provide strong observational evidence that the CGR is coupled to changes in both temperature and water storage at the global scale. The role of water storage is also stronger than can be diagnosed from precipitation⁵ (Extended Data Fig. 5) or precipitation conditional on the ENSO phase⁶. However, these findings differ from the recent results of Jung and colleagues²², who suggested that the global mean net ecosystem exchange (NEE) simulated by statistical models (FluxCom²³)

and physical carbon cycle models (DGVMs⁹) responds to temperature rather than to water storage. To investigate this discrepancy, we reproduce their approach²² (Methods) and find that, while our observations indicate that CGR is highly correlated to global water storage changes (Fig. 4a, circle), modelled NEE fails to reproduce this pattern and is instead mostly correlated to temperature (Fig. 4a, squares). Here, we suggest that this occurs because models underestimate the magnitude of water-driven NEE variations at the global scale (Fig. 4b, c). We find that the water-driven NEE of a given model is (except for one model) directly correlated to its simulated global mean water storage (Fig. 4d, Supplementary Fig. 1). This internal model relationship indicates that a link exists between global mean water storage and its resulting global effect on NEE, which directly supports our observation-based results. We note that this global relationship also holds for the temperature-driven response (Fig. 4d). Therefore, the correlations reported in Fig. 4a for total model NEE are directly controlled by the relative importance of the temperature-driven and water-driven NEE components (Fig. 4b, c). Our observations (Fig. 4a, circle) thus suggest that simulated global NEE may appear to be dominated by temperature effects because the amplitude of water-driven NEE is underestimated. This might indicate that the modelled NEE response is not sensitive enough to soil moisture or it might suggest a role of non-modelled processes that are strongly regulated by other types of water storage changes (for example, the access of deep roots to groundwater²⁴ or the response of inland waters and wetland ecosystems²⁵). In addition, inaccuracies in the precipitation forcing as well as missing water reservoirs in model hydrology might also affect water-driven NEE signals. Compared to GRACE observations, models display a widespread tendency towards underestimating the importance of low-frequency (inter-annual) water storage anomalies and are dominated by short-term fluctuations (Methods, Extended Data Figs. 6–8). This is probably explained by the limited (or absent) representation of deep soil layers, groundwater, wetlands and surface waters, which respond more slowly to climate forcing and have a much longer residence time than root-zone soil moisture. Interestingly, we find that the fraction of IAV (Methods) in modelled water storage imposes a strong upper limit on how much IAV can ultimately be found in modelled water-driven NEE (Extended Data Fig. 9). As a result, the amplitude of water-driven NEE at the interannual timescale (Fig. 4b) is limited by a lack of long-term persistence in the underlying water storage signal.

By partitioning the water storage signal into six land-cover classes (Supplementary Fig. 2), we find that GRACE observations and models agree that semi-arid regions dominate the global mean water storage signal (Extended Data Fig. 10), even though models do not correlate very well with the actual signal observed by GRACE (Supplementary Fig. 3). These findings support recent results suggesting that semi-arid (and thus water-limited) ecosystems are responsible for most of the

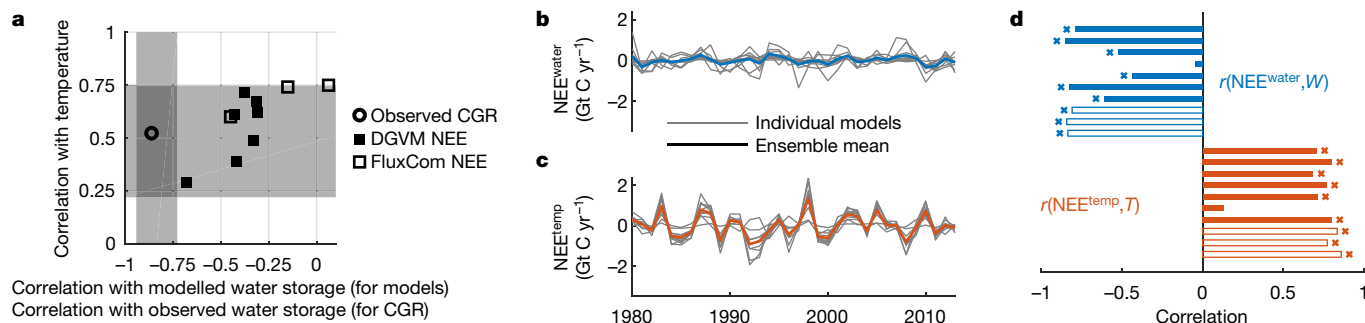


Fig. 4 | Observed and modelled relations between global water storage, temperature and carbon fluxes. **a**, Correlations of the land carbon sink with global mean temperature and global mean water storage (based on simulated soil moisture when correlating with model NEE and based on GRACE when correlating with observed CGR). Grey shading indicates the 95% confidence intervals (from Fig. 2b) for the observed relationships (circle). Solid and hollow squares indicate the relationships obtained with

DGVMs and the FluxCom models, respectively. **b**, **c**, Global means of the NEE signals driven by soil moisture ($\text{NEE}_{\text{water}}$) (**b**) and temperature (NEE_{temp}) (**c**) (Methods). **d**, The $\text{NEE}_{\text{water}}$ of a given model is correlated to its simulated global mean soil moisture (W ; blue bars), and NEE_{temp} is correlated to global mean temperature (T ; orange bars), indicating an internal consistency between the global means of these two climatic drivers and their associated NEE response.

CGR IAV^{26,27}. However, while we find that GRACE water storage in semi-arid regions is well correlated with CGR, our analysis also suggests a possible role of tropical forests (Fig. 1c, d, Supplementary Fig. 4).

In summary, we have provided observational evidence that the IAV of the CO₂ growth rate is tightly coupled to TWS changes. The sensitivities derived here represent the aggregated response of processes that operate at smaller spatial scales²². For this reason, they are not directly transferable to the ecosystem scale but may still provide a valuable metric for evaluating and constraining Earth system models^{2,6,10}. Our results suggest that current models might underestimate the response of ecosystems to global changes in water availability. Models typically respond only to shallow soil moisture and are therefore less sensitive to IAV in water storage. They might also miss the response to changes in non-modelled water reservoirs such as wetlands or surface waters. The presented findings offer new perspectives on the use of satellite observations of water storage for global carbon cycle research. Projections of inter-annual as well as long-term water storage changes from hydrological models still display large uncertainties^{28,29}, and will need to be better assessed in order to reduce uncertainties in projections of future land carbon uptake. As an additional complexity, estimates of future TWS are themselves very dependent on how transpiration will be regulated by vegetation in a world of rising CO₂ concentrations³⁰. Such evidence of the interplay between the water and carbon cycles also highlights the need for stronger interactions between the hydrological and biogeochemical research communities.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0424-4>.

Received: 18 December 2017; Accepted: 14 June 2018;

Published online 29 August 2018.

- Le Quéré, C. et al. Global carbon budget 2017. *Earth Syst. Sci. Data* **10**, 405–448 (2018).
- Cox, P. M. et al. Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494**, 341–344 (2013).
- Wang, W. et al. Variations in atmospheric CO₂ growth rates coupled with tropical temperature. *Proc. Natl Acad. Sci. USA* **110**, 13061–13066 (2013).
- Wang, X. et al. A two-fold increase of carbon cycle sensitivity to tropical temperature variations. *Nature* **506**, 212–215 (2014).
- Wang, J., Zeng, N. & Wang, M. R. Interannual variability of the atmospheric CO₂ growth rate: roles of precipitation and temperature. *Biogeosciences* **13**, 2339–2352 (2016).
- Fang, Y. et al. Global land carbon sink response to temperature and precipitation varies with ENSO phase. *Environ. Res. Lett.* **12**, 064007 (2017).
- Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F. & Watkins, M. M. GRACE measurements of mass variability in the Earth System. *Science* **305**, 503–505 (2004).
- Peters, G. P. et al. Towards real-time verification of CO₂ emissions. *Nat. Clim. Chang.* **7**, 848–850 (2017).
- Sitch, S. et al. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences* **12**, 653–679 (2015).
- Friedlingstein, P. et al. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Clim.* **27**, 511–526 (2014).
- Keeling, C. D., Whorf, T. P., Wahlen, M. & Vanderpligt, J. Interannual extremes in the rate of rise of atmospheric carbon-dioxide since 1980. *Nature* **375**, 666–670 (1995).
- Ciais, P. et al. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature* **437**, 529–533 (2005).
- Phillips, O. L. et al. Drought sensitivity of the Amazon rainforest. *Science* **323**, 1344–1347 (2009).
- Beer, C. et al. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science* **329**, 834–838 (2010).
- Humphrey, V., Gudmundsson, L. & Seneviratne, S. I. A global reconstruction of climate-driven subdecadal water storage variability. *Geophys. Res. Lett.* **44**, 2300–2309 (2017).
- Lucht, W. et al. Climatic control of the high-latitude vegetation greening trend and Pinatubo effect. *Science* **296**, 1687–1689 (2002).
- Trenberth, K. E. & Dai, A. Effects of Mount Pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering. *Geophys. Res. Lett.* **34**, (2007).
- Mercado, L. M. et al. Impact of changes in diffuse radiation on the global land carbon sink. *Nature* **458**, 1014–1017 (2009).
- Döll, P., Müller Schmied, H., Schuh, C., Portmann, F. T. & Eicker, A. Global-scale assessment of groundwater depletion and related groundwater abstractions: combining hydrological modeling with information from well observations and GRACE satellites. *Wat. Resour. Res.* **50**, 5698–5720 (2014).
- Rodell, M. et al. The global land data assimilation system. *Bull. Am. Meteorol. Soc.* **85**, 381–394 (2004).
- Ni, S. et al. Global terrestrial water storage changes and connections to ENSO events. *Surv. Geophys.* **39**, 1–22 (2018).
- Jung, M. et al. Compensatory water effects link yearly global land CO₂ sink changes to temperature. *Nature* **541**, 516–520 (2017).
- Tramontana, G. et al. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **13**, 4291–4313 (2016).
- Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B. & Otero-Casal, C. Hydrologic regulation of plant rooting depth. *Proc. Natl Acad. Sci. USA* **114**, 10572–10577 (2017).
- Battin, T. J. et al. The boundless carbon cycle. *Nat. Geosci.* **2**, 598–600 (2009).
- Poulter, B. et al. Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature* **509**, 600–603 (2014).
- Ahlstrom, A. et al. The dominant role of semi-arid ecosystems in the trend and variability of the land CO₂ sink. *Science* **348**, 895–899 (2015).
- Orlowsky, B. & Seneviratne, S. I. Elusive drought: uncertainty in observed trends and short- and long-term CMIP5 projections. *Hydrol. Earth Syst. Sci.* **17**, 1765–1781 (2013).
- Scanlon, B. R. et al. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Natl Acad. Sci. USA* **115**, E1080–E1089 (2018).
- Swann, A. L. S., Hoffman, F. M., Koven, C. D. & Randerson, J. T. Plant responses to increasing CO₂ reduce estimates of climate impacts on drought severity. *Proc. Natl Acad. Sci. USA* **113**, 10019–10024 (2016).

Acknowledgements All datasets supporting the results of this paper are openly accessible from the references listed in Supplementary Table 1. This research was funded by the European Research Council DROUGHT-HEAT project (contract 617518). P.C. was supported by the European Research Council Synergy grant ERC-2013-SyG-610028 IMBALANCE-P. We thank M. Jung and U. Weber for providing the water availability index used in FluxCom and R. Wartenburger for technical support. We gratefully thank the following data providers and model developers for their continuous efforts and for sharing their data: the NASA Jet Propulsion Laboratory, the NOAA Earth System Research Laboratory, the Global Carbon Project, the WaterGAP Global Hydrology Model (WGHM), the Global Land Data Assimilation System (GLDAS), Multi-Source Weighted-Ensemble Precipitation (MSWEP), the Global Precipitation Climatology Project (GPCP), the University of East Anglia Climatic Research Unit (CRU), Berkeley Earth, and all contributors as well as data providers to the FluxCom initiative and the TRENDY experiment version 3, which included the models CABLE, CLM, ISAM, JSBACH, JULES, LPJ, LPJ-GUESS, LPX-Bern, ORCHIDEE, VEGAS and VISIT.

Reviewer information Nature thanks A. Dolman, C. Funk and B. Zaitchik for their contribution to the peer review of this work.

Author contributions V.H., S.I.S., J.Z. and P.C. designed the study. V.H. conducted the data analysis with support from L.G., J.Z., S.S. and S.I.S., and wrote the manuscript. The interpretation, final text and figures resulted from the contributions of all co-authors.

Competing interests : The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0424-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0424-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to V.H. or S.I.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

GRACE water mass changes. From 2002 to 2017, the twin GRACE satellites have measured monthly anomalies of the Earth's gravity field^{47,31} that can be used to retrieve relative changes in water storage, both on land and in the ocean, at a spatial resolution of about 300 km^{32,33}. Over land, these observations reflect net changes in TWS, which include groundwater, soil moisture, surface waters, snow, land ice and water stored in the biosphere. These observations of water mass redistribution are consistent with other observed geophysical constraints such as changes in sea level³⁴ and polar motion³⁵ and correlate with satellite observations of surface soil moisture³⁶ as well as with changes in precipitation and temperature³⁷. Here, we use the Jet Propulsion Laboratory mass concentration (mascon) solution^{38,39} and exclude the contribution of Greenland and Antarctica in order to obtain a global mean TWS signal for all available months between April 2002 and December 2016. We isolate the IAV by subtracting the mean seasonal cycle and remove the linear trend using simple linear regression. Because we focus on global and regional averages over very large spatial domains, using different GRACE solutions (Supplementary Table 2) has very little impact (Supplementary Fig. 5). A comprehensive comparison can be found for example in ref. ⁴⁰. We recommend checking different solutions for local case studies. The GRACE Follow-On satellites, which were launched in May 2018, will replace the GRACE satellites and are expected to extend the gravity record by another 5–10 years.

Derivation of the CGR. We use monthly time series of atmospheric CO₂ concentration from the Greenhouse Gas Marine Boundary Layer Reference of the National Oceanic and Atmospheric Administration (NOAA/ESRL)^{41,42}. This dataset compiles measurements of weekly air samples from the Cooperative Global Air Sampling Network since 1980. Like ref. ³, we derive monthly CGR as the first-order difference of CO₂ concentrations between two successive months. We then remove the mean seasonal cycle and apply a 12-month moving sum to convert monthly values into annual CGR. For completeness, we also repeat the analysis at the yearly scale using estimates of the Residual Land Sink from the Global Carbon Project⁴, and show that this does not affect the conclusions of the paper (Supplementary Figs. 6 and 7).

GRACE-REC. GRACE-REC is a statistical reconstruction of GRACE. The statistical approach used to generate the reconstruction of past TWS anomalies is explained in detail in ref. ¹⁵. In summary, a statistical model forced with daily precipitation and temperature anomalies is trained with GRACE observations and used to reconstruct past changes in water storage. In ref. ¹⁵, the precipitation forcing is based on the meteorological re-analysis of the European Centre for Medium-Range Weather Forecasts (ERA-Interim), which has some limitations in representing tropical precipitation compared to other datasets. In this study, we reconstruct past TWS anomalies with the same approach but using a recently published merged daily precipitation product⁴³. Using this new precipitation dataset leads to a small improvement in model performance, but there may still be limitations in the accuracy of the precipitation data, in particular over tropical regions. This updated TWS reconstruction is publicly accessible as part of this publication (Supplementary Table 1). **Global and regional land averages.** The contribution of Greenland and Antarctica is removed for all analyses. Global and regional averages are weighted according to the land area of grid cells. The tropical domain definition used in this paper ranges from 24° S to 24° N, as in ref. ³. Information on the datasets^{44–49} used to generate land averages can be obtained from Supplementary Table 1. Land cover classes are based on MODIS MCD12C1 (Supplementary Fig. 2).

Monte Carlo estimate of correlation significance and uncertainty intervals. We estimate the 95% confidence interval of correlation coefficients (such as the confidence intervals reported in Fig. 2), null hypothesis distributions for two-tailed significance testing, as well as distributions for univariate and bivariate linear sensitivities in Fig. 3c and d using moving-block bootstrapping⁵⁰. The selection of the block length is a compromise between accounting for the effect of autocorrelation in the time series and keeping a sufficiently large block sample size so that random resamples stay independent. Using different block length selectors⁵⁰, we defined a block length of 12 months for monthly analyses (block length = 12). For yearly analyses, the block length was defined as 1 year (block length = 1, which is equivalent to a simple bootstrap approach), because the autocorrelation of time series was not significant at the yearly scale. The same procedure was applied to all datasets considered with 10,000 random resamples.

Time intervals for various datasets and exclusion of years 1991–1993. The correlations reported in Figs. 2 and 4 are computed over heterogeneous time intervals to make use of as much data as is currently available (Supplementary Table 3). Pairs of time series are de-trended over their common time interval. To assess these correlations over a time period as homogeneous as possible, we repeat the analysis for these figures over the period 2002–2013 only and find that our conclusions remain unchanged (Supplementary Figs. 8, 9). The eruption of Mt Pinatubo strongly affected radiation budgets, which perturbed both CGR^{16,18} and the water cycle¹⁷, explaining the de-coupling between CGR and water storage changes. We also reproduce Fig. 2 without discarding the years following the eruption of Mt Pinatubo (1991–1993) and find that our main conclusions remain unchanged

(Supplementary Fig. 10), although the correlation between CGR and water-related variables decreases (as can be expected from Fig. 1a).

Approach of ref. ²² to separate temperature-driven and water-driven NEE signals. In a recent paper, Jung and colleagues²² performed a global analysis of the drivers of NEE IAV using DGVMs and statistical models trained on flux tower measurements. They estimated the sensitivity of NEE IAV to climate drivers by fitting local multivariate regressions to the model outputs. With this approach, the simulated soil moisture and the observed temperature forcings are used to fit a linear statistical model of the monthly carbon flux response calculated by the more complex DGVMs and upscaling models. We replicated the analysis performed in ref. ²² on DGVMs using the same set of seven DGVMs (from TRENDY v3/S2)⁹ as well as upscaling models (FluxCom²³).

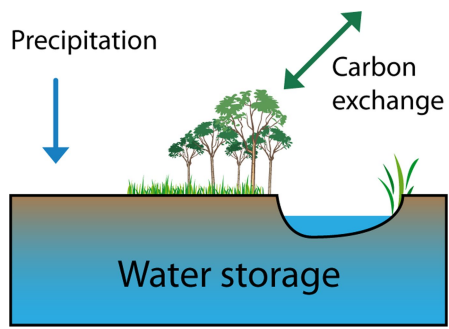
Fraction of IAV. The fraction of IAV quantifies the importance of low frequency variability in the overall variance of a given signal. It is computed as:

$$F_{IAV} = \frac{\text{Var}(X_{\text{yearly}})}{\text{Var}(X_{\text{monthly}})}$$

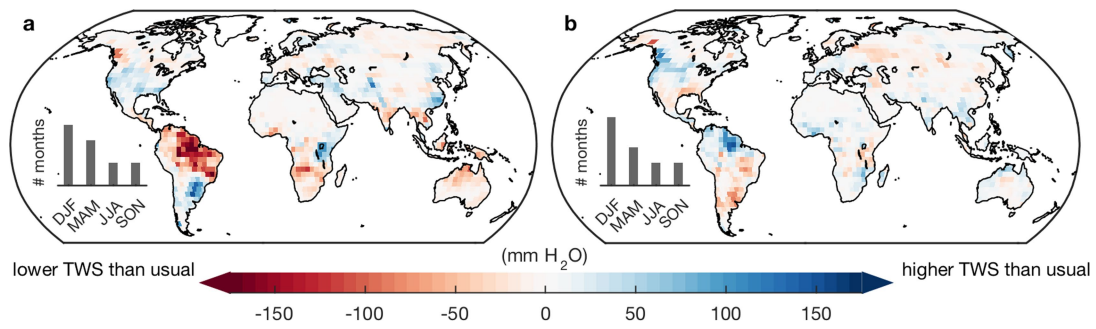
where, Var() denotes the variance estimator, X_{monthly} is the de-seasonalized and de-trended monthly time series and X_{yearly} is the yearly time series (computed from X_{monthly}). This indicator is also illustrated in Extended Data Fig. 6. Although it is much simpler in practice, this approach shares the same idea as analysing the relative importance of low frequencies in a signal's power spectrum. When adding GRACE measurement errors to the GRACE-REC estimates of IAV fraction in Extended Data Fig. 8, we use the measurement errors provided with the original JPL mascons, but without applying a conservative scale factor of 2 to the diagonal elements of the formal covariance matrix (see ref. ³⁸).

Data availability. All datasets generated or analysed during this study are available from the links listed in Supplementary Table 1. The source data for Figs. 1a–b, 2, 3 and 4 are additionally provided as spreadsheets with the online version of the paper.

- Wahr, J., Swenson, S., Zlotnicki, V. & Velicogna, I. Time-variable gravity from GRACE: first results. *Geophys. Res. Lett.* **31**, L11501 (2004).
- Wahr, J., Molenaar, M. & Bryan, F. Time variability of the Earth's gravity field: hydrological and oceanic effects and their possible detection using GRACE. *J. Geophys. Res. Solid Earth* **103**, 30205–30229 (1998).
- Wouters, B. et al. GRACE, time-varying gravity, Earth system dynamics and climate change. *Rep. Prog. Phys.* **77**, (2014).
- Cazenave, A. et al. The rate of sea-level rise. *Nat. Clim. Chang.* **4**, 358–361 (2014).
- Adhikari, S. & Ivins, E. R. Climate-driven polar motion: 2003–2015. *Sci. Adv.* **2**, (2016).
- Abelen, S. & Seitz, F. Relating satellite gravimetry data to global soil moisture products via data harmonization and correlation analysis. *Remote Sens. Environ.* **136**, 89–98 (2013).
- Humphrey, V., Gudmundsson, L. & Seneviratne, S. I. Assessing global water storage variability from GRACE: trends, seasonal cycle, subseasonal anomalies and extremes. *Surv. Geophys.* **37**, 357–395 (2016).
- Wiese, D. N., Landerer, F. W. & Watkins, M. M. Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Wat. Resour. Res.* **52**, 7490–7502 (2016).
- Watkins, M. M., Wiese, D. N., Yuan, D. N., Boening, C. & Landerer, F. W. Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *J. Geophys. Res. Solid Earth* **120**, 2648–2671 (2015).
- Scanlon, B. R. et al. Global evaluation of new GRACE mascon products for hydrologic applications. *Wat. Resour. Res.* **52**, 9412–9429 (2016).
- Masarie, K. A. & Tans, P. P. Extension and integration of atmospheric carbon-dioxide data into a globally consistent measurement record. *J. Geophys. Res. D* **100**, 11593–11610 (1995).
- Dlugokencky, E. & Tans, P. Trends in Atmospheric Carbon Dioxide. <http://www.esrl.noaa.gov/gmd/ccgg/trends/> (National Oceanic and Atmospheric Administration, Earth System Research Laboratory (NOAA/ESRL), 2014).
- Beck, H. E. et al. MSWEP: 3-hourly 0.25 degrees global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.* **21**, 589–615 (2017).
- Adler, R. F. et al. The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol.* **4**, 1147–1167 (2003).
- Huffman, G. J., Adler, R. F., Bolvin, D. T. & Gu, G. J. Improving the global precipitation record: GPCP version 2.1. *Geophys. Res. Lett.* **36**, (2009).
- Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
- Rohde, R. et al. A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinform. Geostatist.* **01**, <http://doi.org/10.4172/2327-4581.1000101> (2013).
- Wolter, K. & Timlin, M. S. Measuring the strength of ENSO events: how does 1997/98 rank? *Weather* **53**, 315–324 (1998).
- Friedl, M. A. et al. MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **114**, 168–182 (2010).
- Mudelsee, M. *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods* 2nd edn, Chs 4 and 7 (Springer, Cham, 2014).

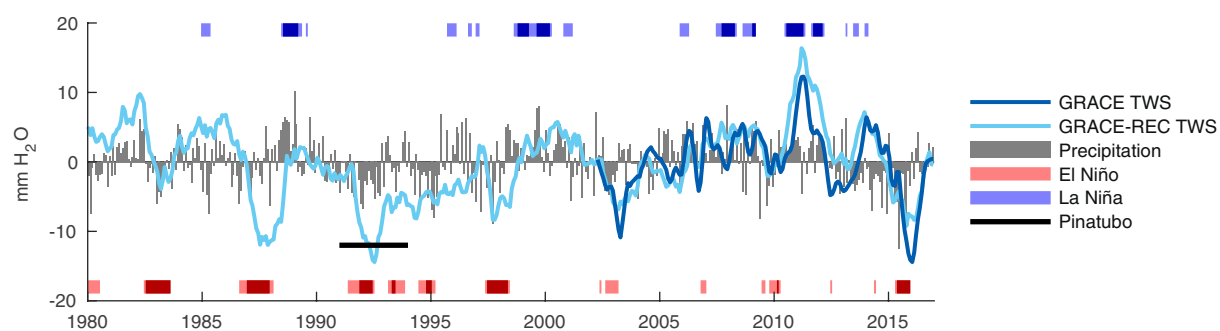


Extended Data Fig. 1 | Ecosystems respond to water storage. Water storage is more relevant than precipitation when investigating the impacts of changes in water availability on ecosystems.



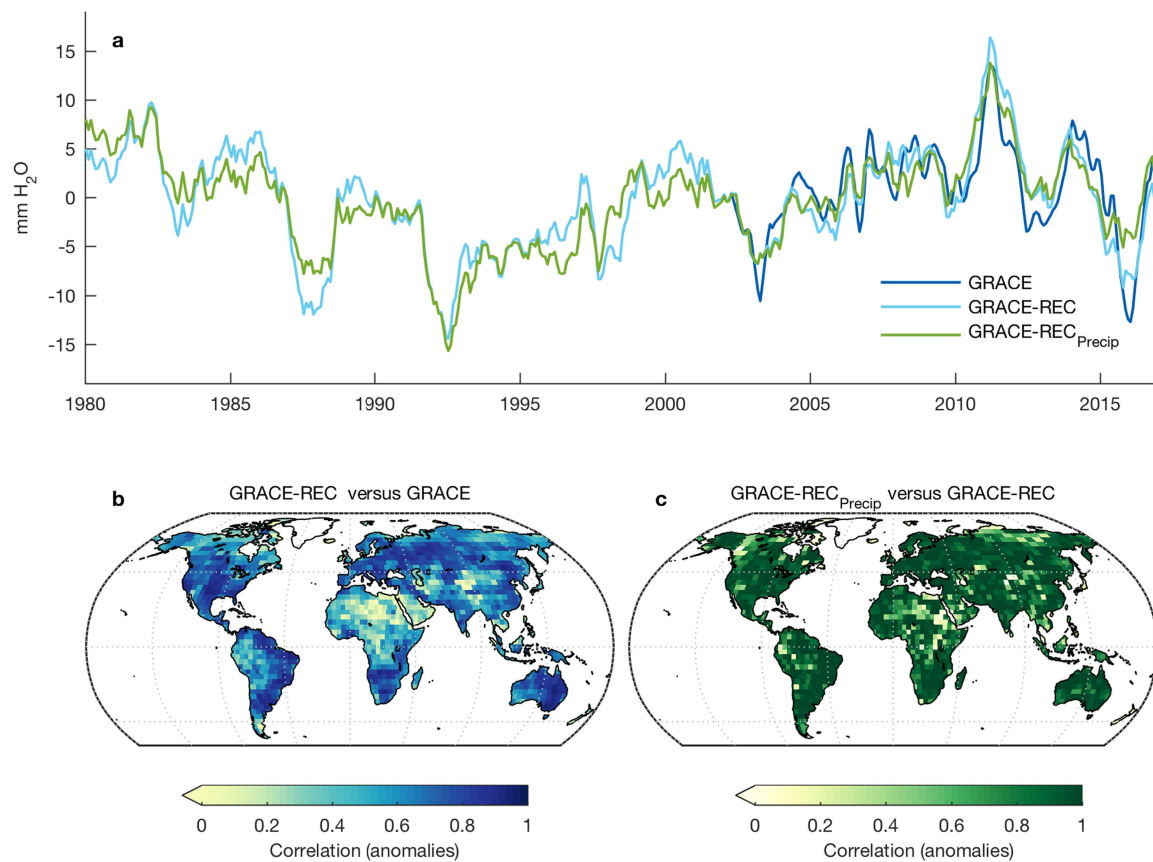
Extended Data Fig. 2 | Reproduction of Fig. 1c, d with GRACE-REC. Composite mean TWS anomalies associated with the 5% highest (a) and 5% lowest (b) monthly CGR ($n = 20$ months in each case) based on

GRACE-REC (that is, covering the 1980–2016 time period). Inset bar plots indicate the season of the months used in the composite.



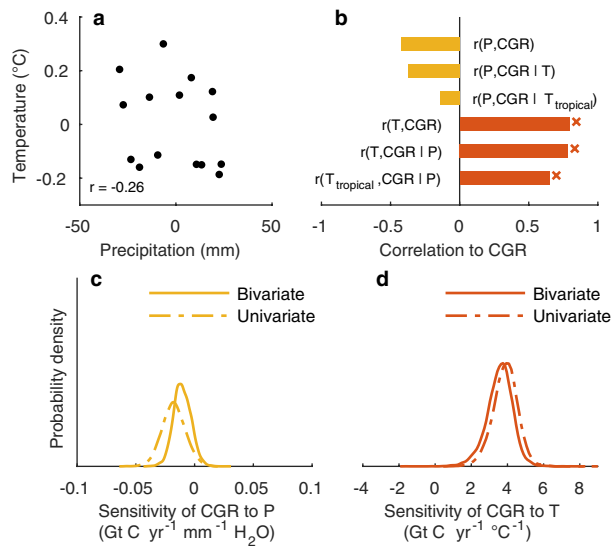
Extended Data Fig. 3 | ENSO, precipitation and TWS. Because it integrates precipitation anomalies, water storage is slightly phase-shifted with respect to ENSO and precipitation time series. Here, El Niño

(La Niña) conditions correspond to the periods where the Multivariate ENSO Index (MEI) exceeds 0.5 (−0.5). The strongest ENSO events (MEI > 1 or MEI < −1) are shown in darker colours.



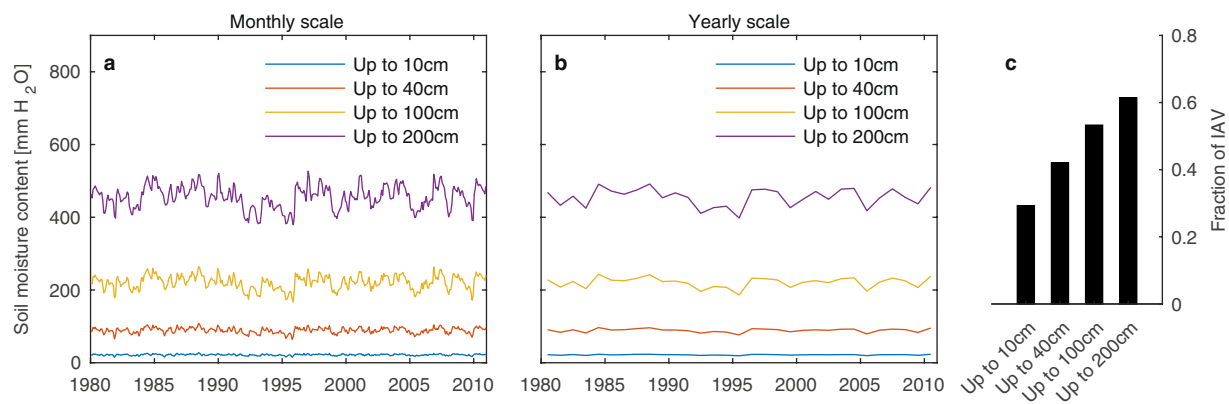
Extended Data Fig. 4 | Dominant contribution of precipitation to TWS anomalies. **a**, Global means of GRACE, GRACE-REC and GRACE-REC driven only with precipitation anomalies. The statistical reconstruction of GRACE (GRACE-REC) is calibrated with both precipitation and temperature information¹⁵. We use this model to predict the precipitation-driven component of the TWS signal (by setting temperature variability

to zero). Most of the global TWS signal can be reconstructed based on precipitation anomalies only. **b**, Performance of the GRACE-REC model at the grid scale. **c**, Contribution of precipitation to the locally reconstructed TWS. A comparison between GRACE-REC, global hydrological models and GRACE can also be found in ref. ¹⁵.



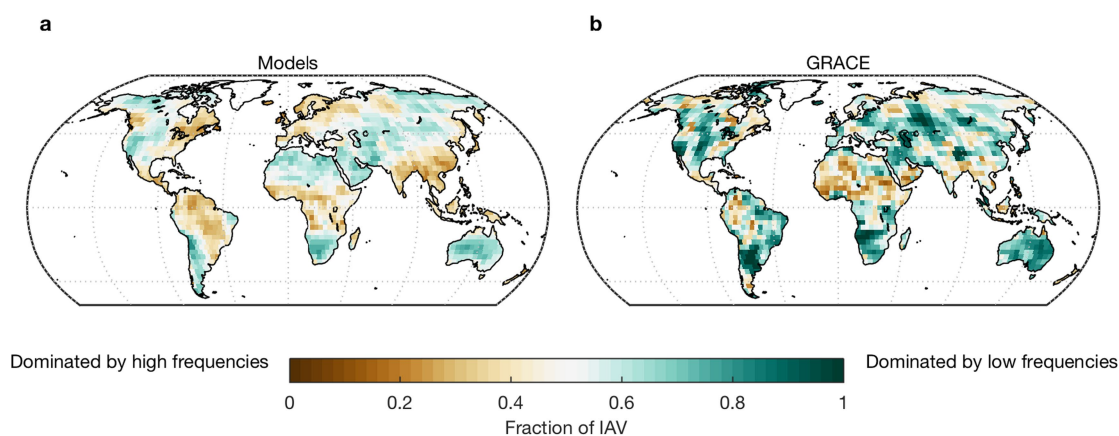
Extended Data Fig. 5 | Reproduction of Fig. 3 with mean precipitation.

Same as Fig. 3, but using yearly precipitation P from the Global Precipitation Climatology Project (with a 4-month lag) instead of TWS from GRACE. Significance ($P < 0.05$, $n = 15$; Methods) is indicated with crosses.



Extended Data Fig. 6 | Illustration of soil moisture signals with different fractions of IAV. The fraction of IAV quantifies the importance of low frequency variability in the overall variance of a given signal. Here, it is defined as the ratio between the variance of the yearly (de-trended) time series (b) and the variance of the monthly anomalies (a)

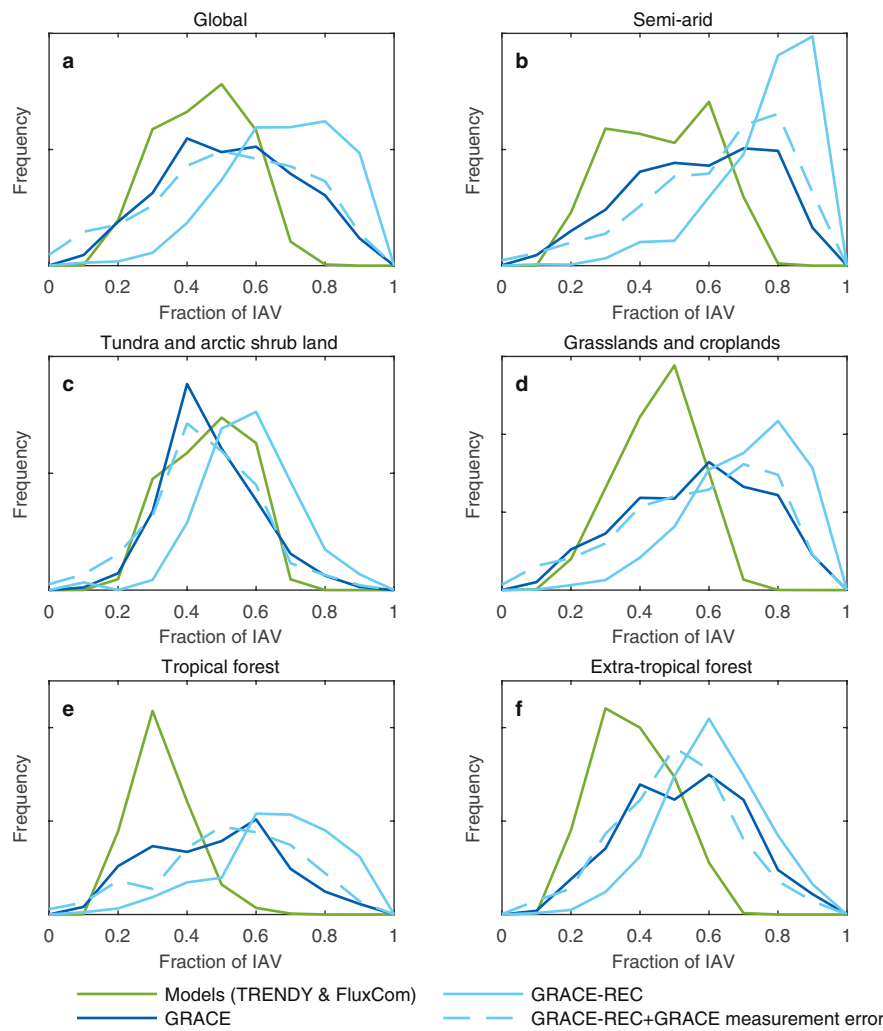
(see Methods). The fraction of IAV tends to increase when deeper soil layers are included (c). This is because deeper layers have a longer residence time (or memory) and thus respond more slowly to changes in the meteorological forcing. Illustrative data based on GLDAS2-Noah, extracted for Spain (4.25° W, 40.25° N).



Extended Data Fig. 7 | Fraction of IAV in water storage changes.

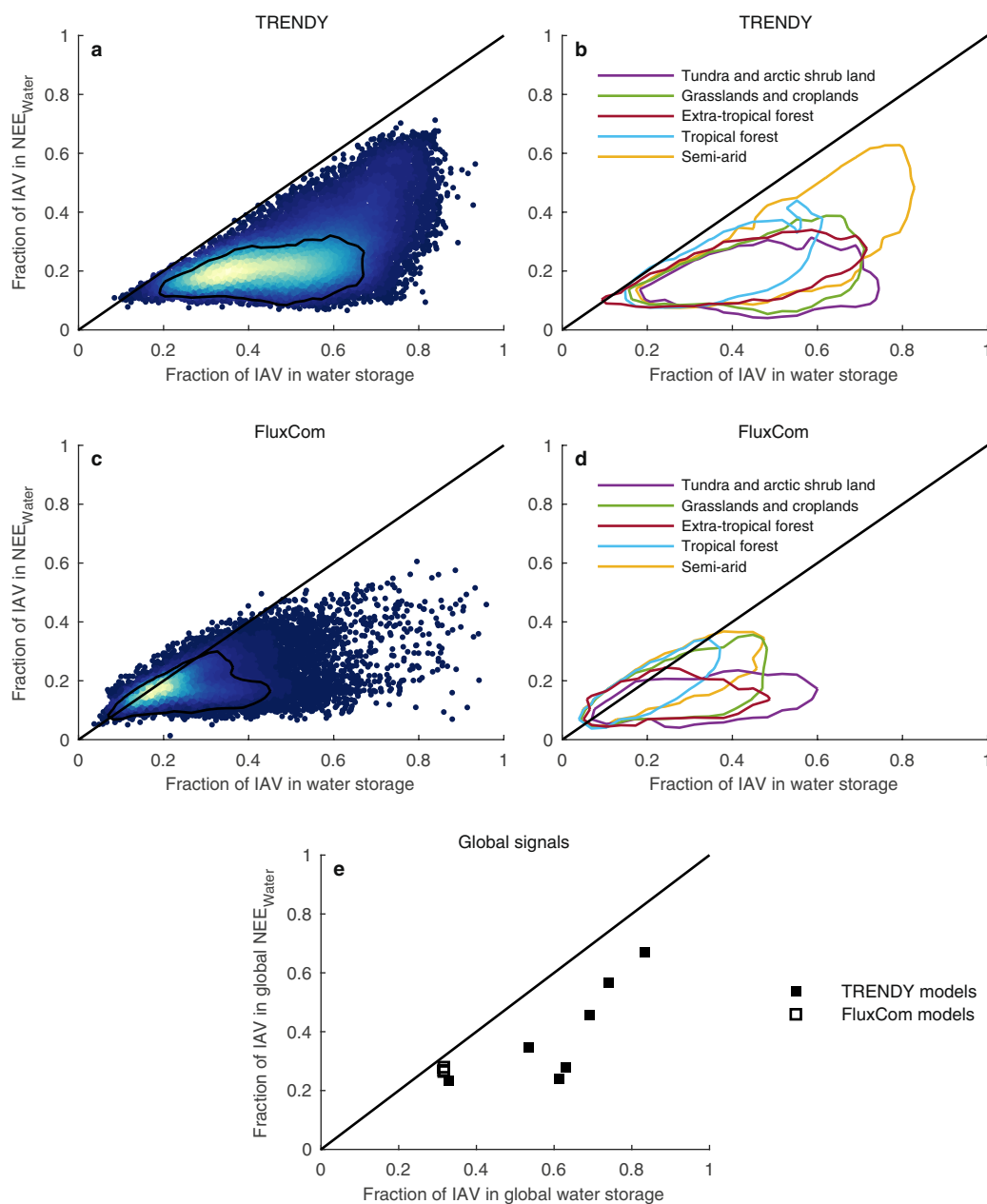
a, Average fraction of IAV in water storage changes simulated by DGVMs and FluxCom (which typically only include root-zone soil moisture).
b, Fraction of IAV in water storage changes observed by GRACE (which include all water reservoirs). To ensure comparability between models and GRACE, model outputs were first averaged to the spatial resolution of

GRACE. We note that unlike modelled soil moisture, GRACE observations suffer from measurement errors that tend to increase the high-frequency (month-to-month) variability. Therefore, the fraction of IAV retrieved from GRACE would be even higher if there were no measurement errors in GRACE.



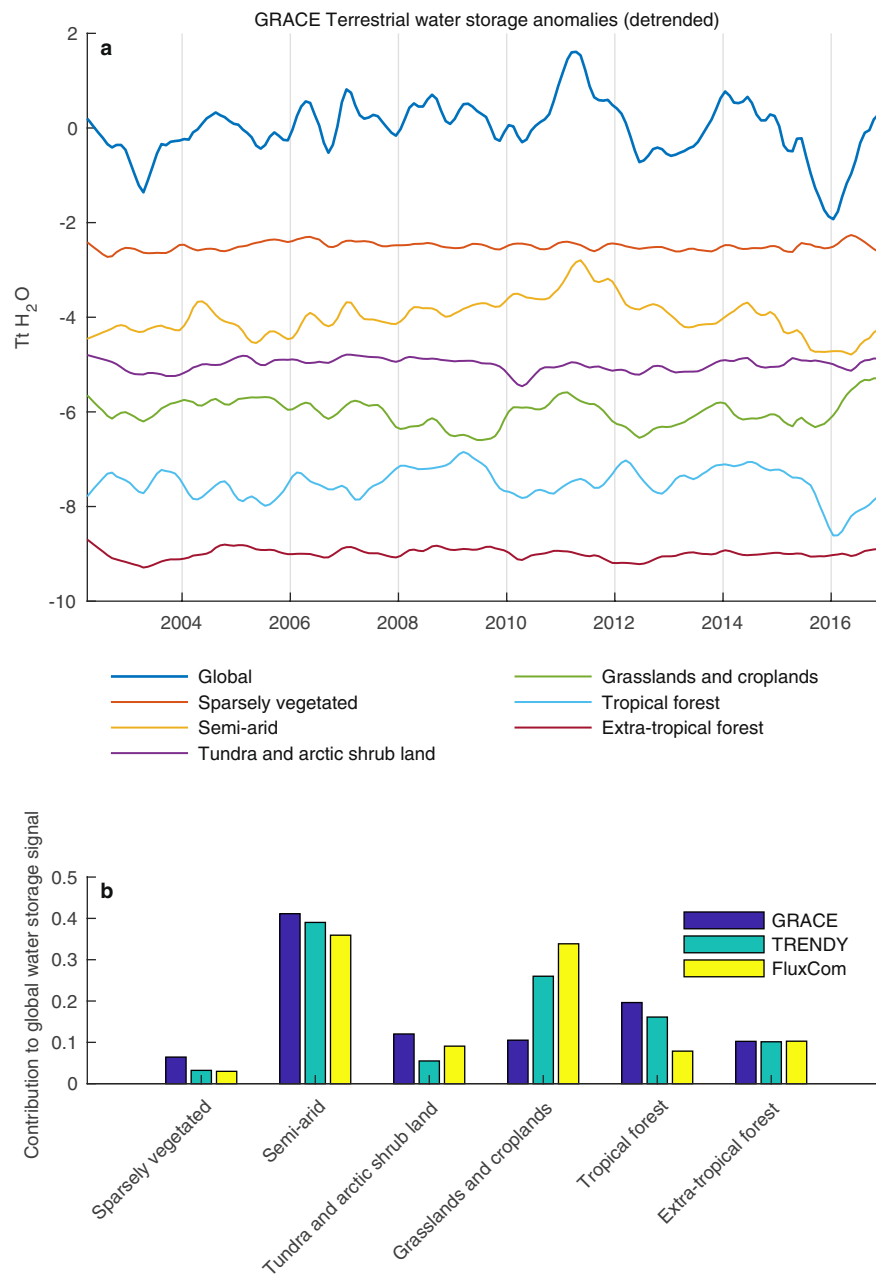
Extended Data Fig. 8 | Distribution of the fraction of IAV by land cover classes. This compares the values shown in the maps of Extended Data Fig. 7 for different land-cover classes. The fraction of IAV found in GRACE TWS (dark blue) is higher compared to models (green). Because GRACE observations are contaminated by high-frequency measurement errors, the fraction of IAV found in GRACE is shifted towards lower

values. Here, the fraction of IAV derived from GRACE-REC (light blue) may provide a more robust estimate of the actual fraction of IAV in TWS. Adding GRACE measurement errors (as provided with GRACE NASA-JPL data) to the GRACE-REC data reproduces very well the overall shift (dashed light blue) towards lower values that occurs with original GRACE data.



Extended Data Fig. 9 | Relationship between the fraction of IAV in model water storage and the fraction of IAV in NEE_{water} . **a, c**, Mean fraction of IAV obtained at all grid cells for TRENDY (**a**) and FluxCom (**c**), with point cloud density indicated by the colour shading. The fraction of IAV in NEE_{water} is directly limited by the fraction of IAV present in the

underlying water storage signal. **b, d**, The same as **a** and **c**, stratified by land-cover class. In land-cover classes that are typically moisture-limited (for example, semi-arid), the fraction of IAV in NEE_{water} is potentially strongly limited by the fraction of IAV in water storage. **e**, This relationship is also found for the global mean signals of the individual models.



Extended Data Fig. 10 | Contribution of six different land cover types to the global water storage signal. a, GRACE TWS anomalies by land-cover type, smoothed with a 6-month moving average and offset for readability. **b,** Regional contributions to the global water storage signal. High values

indicate that a region bears a high contribution to the overall global mean water storage signal. This metric is based on the definition proposed in ref.²⁷ for analysing regional contributions to global net biome production. The value reported for the models is the mean across all models.

Deep learning of aftershock patterns following large earthquakes

Phoebe M. R. DeVries^{1,2*}, Fernanda Viégas³, Martin Wattenberg³ & Brendan J. Meade¹

Aftershocks are a response to changes in stress generated by large earthquakes and represent the most common observations of the triggering of earthquakes. The maximum magnitude of aftershocks and their temporal decay are well described by empirical laws (such as Bath's law¹ and Omori's law²), but explaining and forecasting the spatial distribution of aftershocks is more difficult. Coulomb failure stress change³ is perhaps the most widely used criterion to explain the spatial distributions of aftershocks^{4–8}, but its applicability has been disputed^{9–11}. Here we use a deep-learning approach to identify a static-stress-based criterion that forecasts aftershock locations without prior assumptions about fault orientation. We show that a neural network trained on more than 131,000 mainshock–aftershock pairs can predict the locations of aftershocks in an independent test dataset of more than 30,000 mainshock–aftershock pairs more accurately (area under curve of 0.849) than can classic Coulomb failure stress change (area under curve of 0.583). We find that the learned aftershock pattern is physically interpretable: the maximum change in shear stress, the von Mises yield criterion (a scaled version of the second invariant of the deviatoric stress-change tensor) and the sum of the absolute values of the independent components of the stress-change tensor each explain more than 98 per cent of the variance in the neural-network prediction. This machine-learning-driven insight provides improved forecasts of aftershock locations and identifies physical quantities that may control earthquake triggering during the most active part of the seismic cycle.

The deep-learning aftershock location forecasts that we have developed are trained and tested using co-seismic slip distributions from the SRCMOD online database of finite-fault rupture models (<http://equake-rc.info/SRCMOD/>). We calculated elastic stress-change tensors for 199 of the SRCMOD slip distributions (118 distinct mainshocks; Supplementary Table 1) at the centroids of 5 km × 5 km × 5 km cells in a volume extending 100 km horizontally from each mainshock rupture plane and 50 km vertically¹². The aftershocks that occurred between one second and one year after the mainshocks in each grid cell (162,741 aftershocks in total) were compiled from the International Seismological Center (ISC) event catalogue. By discretizing the volume around each mainshock in this way, aftershock forecasting can be formulated as a large-scale binary classification problem, with the goal of accurately classifying each 5 km × 5 km × 5 km grid cell in the volume around each mainshock as either ‘containing aftershocks’ or ‘not containing aftershocks’.

Neural networks are machine-learning algorithms that are well suited and widely used to classify data¹³. The neural networks used here are fully connected and have six hidden layers with 50 neurons each and hyperbolic tangent activation functions (13,451 weights and biases in total). The first layer corresponds to the inputs to the neural network; in this case, these inputs are the magnitudes of the six independent components of the co-seismically generated static elastic stress-change tensor calculated at the centroid of a grid cell and their negative values. In neural networks designed for binary classification problems, the final layer is often a single sigmoid. In our case, the output of this final

neuron may be interpreted as the predicted probability that a grid cell generates one or more aftershocks.

The stress changes and aftershock locations associated with about 75% of randomly selected distinct mainshocks were used as training data; the remaining 25% were reserved to test the trained neural networks. The training and testing datasets both consist of the elements of the stress-change tensor as features and the corresponding labels of either 0, for grid cells without aftershocks, or 1, for grid cells with aftershocks.

We assess the accuracy of the neural-network aftershock location forecasts on the test dataset using receiver operating characteristic (ROC) analysis. ROC curves are widely used to assess the efficacy of diagnostic medical tests. To build these curves, the true positive rate of a binary classifier is plotted against the false positive rate for all possible thresholds of the classifier (see Methods for more details). The area under an ROC curve (AUC) then quantifies the overall performance of a test across all thresholds (Fig. 1). The ROC analysis reveals that the neural-network forecast can explain aftershock locations better than can widely used metrics: the merged AUC value across all slip distributions and grid cells in the test dataset for the neural-network forecast is 0.849, which is larger than that of the classic Coulomb failure stress criterion³ (AUC = 0.583) resolved on receiver planes parallel to the average orientation of the mainshock fault ($\Delta\text{CFS}(\mu = 0.4)$, in which μ is the effective coefficient of friction). Neither classifier has particularly high precision, defined as the percentage of grid cells predicted to be positive that actually are positive: the overall precision associated with $\Delta\text{CFS}(\mu = 0.4)$ at a cut-off threshold of 0.01 MPa (Methods) is 3% and that of the neural-network classifier at a threshold of 0.5 is 6% (Fig. 2a, d). Permutation tests (Methods) reveal that the neural-network forecast is significantly better than random assignment for most of the slip distributions in the test dataset: the mean empirical P value is 0.026 across all 57 distributions, and only four distributions are associated with empirical P values larger than 0.1. Additional tests, based on realizations of the training and test datasets that incorporate only one slip distribution per mainshock and variable limits on grid-cell depth depending on the depth of each slip distribution are included in Methods.

The spatial pattern of the deep-learning location forecast can be visualized for the idealized synthetic reference case of an earthquake with a uniform 1 m of slip on a 60-km-long right-lateral strike-slip fault (moment magnitude $M_w \approx 7.0$, Fig. 2; see Extended Data Fig. 1 for an idealized dip-slip case). A location forecast based on the Coulomb failure stress criterion³ for this idealized strike-slip fault would assign a low risk of aftershocks adjacent to the mainshock rupture plane and a heightened risk in lobes extending from the termini of the mainshock rupture plane (Fig. 2e). By contrast, the learned forecast developed here suggests that aftershock risk may be heightened within 10 km of the mainshock fault in all directions (Fig. 2h). This deep-learning forecast is therefore not consistent with the idea of well-defined stress shadows⁶ immediately adjacent to the mainshock.

The learned forecast (Fig. 2h) has implications for the physics of aftershock triggering and earthquake generation. Qualitatively, the

¹Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA. ²Center for Integrative Geosciences and Department of Physics, University of Connecticut, Storrs, CT, USA.

³Google, Cambridge, MA, USA. *e-mail: phoebe.devries@uconn.edu

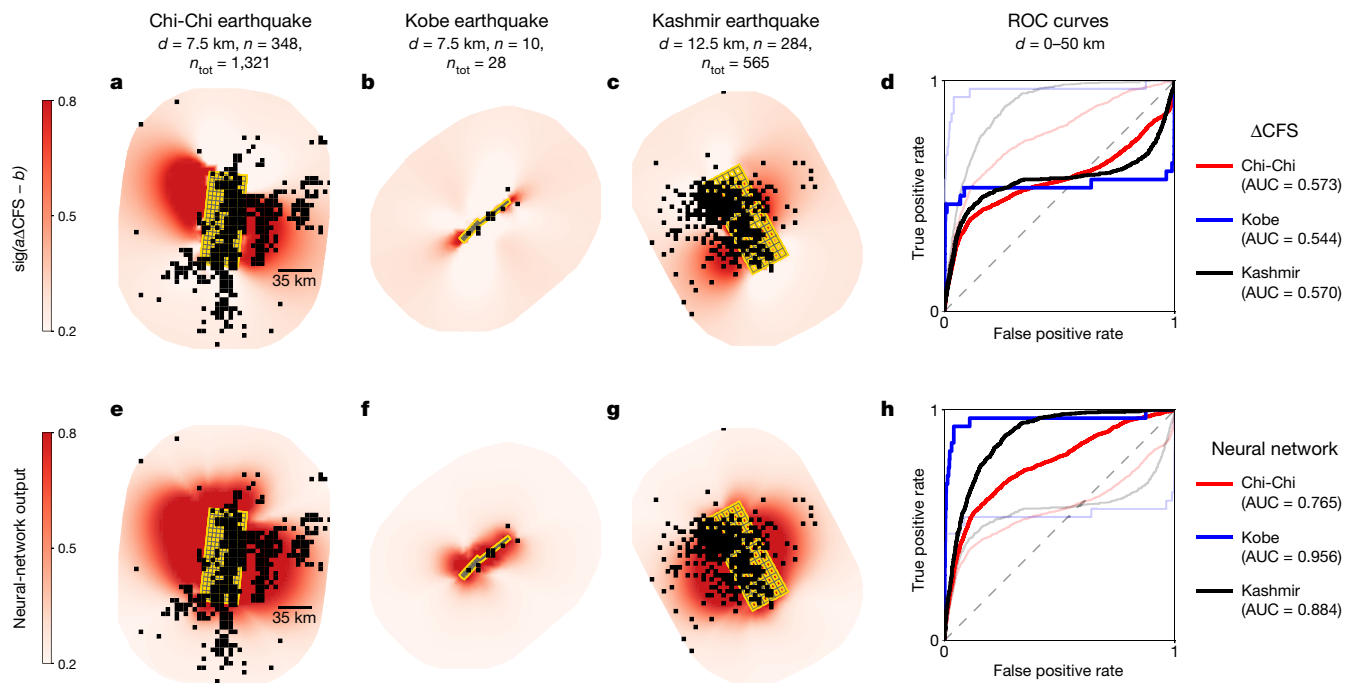


Fig. 1 | Mainshock–aftershock examples. **a–d**, Spatial patterns of $\Delta\text{CFS}(\mu = 0.4)$ for the 1999 $M_w = 7.7$ Chi-Chi earthquake¹⁷ at a depth of 7.5 km (**a**), the 1995 $M_w = 6.9$ Kobe earthquake¹⁸ at a depth of 7.5 km (**b**) and the 2005 $M_w = 7.6$ Kashmir earthquake¹⁹ at a depth of 12.5 km (**c**), along with ROC curves for all three earthquakes across all depths (**d**). In **a–c**, n refers to the number of positive grid cells at the depth shown and n_{tot} is the number of positive grid cells across all depths. A 1:1 grey dashed line is included in **d** for reference. Because of possible sign ambiguities, we calculate four versions of $\Delta\text{CFS}(\mu = 0.4)$ and use the best-performing

sign convention for each slip distribution. In **a–c**, $\Delta\text{CFS}(\mu = 0.4)$ values (in megapascals) are fed through a sigmoid filter $\text{sig}(x) = 1/(1 + e^{-x})$ ($\text{sig}(a\Delta\text{CFS}(\mu = 0.4) - b)$, with $a = 10$, $b = 1$; colour scale) to facilitate comparison to the neural network; faults are shown in yellow and grid cells that contain aftershocks are shown in black. **e–h**, Analogous to **a–d** but for the neural network. To facilitate easy comparison, the ROC curves in **d** are plotted as pale lines in **h** and the ROC curves in **h** are plotted as pale lines in **d**.

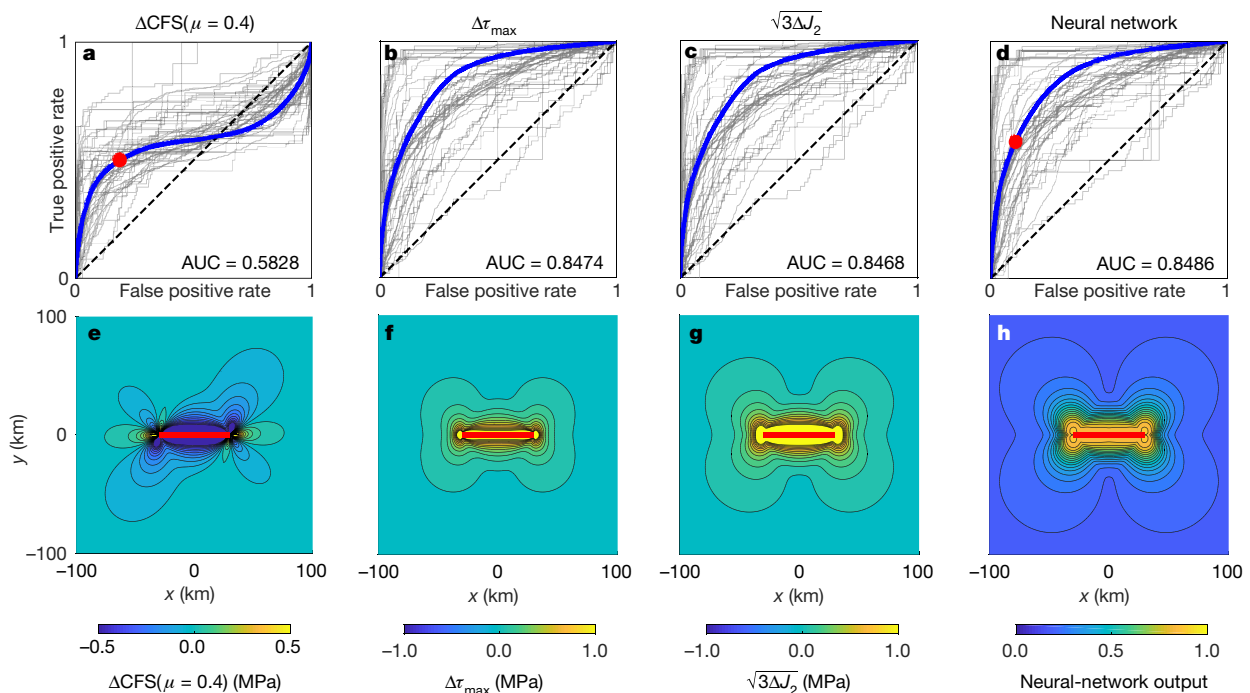


Fig. 2 | Comparison of performance. **a–d**, ROC curves for every slip distribution in the test dataset (grey curves) for $\Delta\text{CFS}(\mu = 0.4)$ (**a**), $\Delta\tau_{\text{max}}$ (**b**), $\sqrt{3}\Delta J_2$ (**c**) and the neural network (**d**). Merged ROC curves are shown in blue and the associated AUC values are listed. The red circles in **a** and **d** highlight the thresholds of 0.01 MPa and 0.5, respectively. **e–h**, For a

synthetic case of a 60-km-long, right-lateral strike-slip fault (red lines) at a depth of 10 km, we show a comparison of the spatial patterns of $\Delta\text{CFS}(\mu = 0.4)$ (**e**), $\Delta\tau_{\text{max}}$ (**f**), $\sqrt{3}\Delta J_2$ (**g**) and the neural network (**h**), averaged over the fault strike.

strike-averaged neural-network forecast appears to closely resemble the spatial patterns of the maximum change in shear stress ($\Delta\tau_{\max}$) and the von Mises yield criterion ($\sqrt{3\Delta J_2}$, in which ΔJ_2 is the second invariant of the deviatoric stress-change tensor; Fig. 2). To examine these potential links to physical quantities quantitatively, we compare a suite of scalar static-stress metrics (including the invariants of the stress-change tensor, Coulomb failure stress change and maximum shear stress change; see Methods)—after they are scaled, shifted and normalized with a sigmoid filter—to the neural-network forecast (Fig. 2h). In addition to Coulomb failure stress change, several of the quantities, including shear stress changes and the invariants of the stress change tensor, have been proposed and used successfully in previous studies of aftershock patterns^{3,14–16}. Of the metrics considered, the maximum change in shear stress, the von Mises yield criterion and the sum of the absolute values of the six independent components of the stress change tensor can explain the largest percentages (more than 98%, or $R^2 > 0.98$ for both the idealized strike-slip and dip-slip cases) of the variance in the strike-averaged learned forecast (see, for example, Fig. 2; Extended Data Fig. 1) within a 300 km \times 300 km area centred on the fault. The last quantity is particularly interesting because it is not invariant under rotation and to our knowledge has not previously been proposed as a quantity that could explain aftershock location patterns. A few other quantities, such as the sum of the absolute values of the shear stress changes on fault-parallel receiver planes can also explain more than 90% of the variance in the strike-slip neural-network forecast (Extended Data Table 1). In other words, without any assumptions about receiver plane orientation or geometry, the neural network identified an aftershock location forecast that is strongly correlated with a small number of physical quantities, most notably the sum of the absolute values of the six independent components of the stress-change tensor, the von Mises yield criterion and the maximum change in shear stress. These results highlight how deep-learning approaches can lead to improved aftershock forecasts and provide physical insights into the mechanisms of earthquake triggering.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0438-y>.

Received: 11 November 2017; Accepted: 10 July 2018;

Published online 29 August 2018.

1. Báth, M. Lateral inhomogeneities of the upper mantle. *Tectonophysics* **2**, 483–514 (1965).
2. Utsu, T. A statistical study on the occurrence of aftershocks. *Geophys. Mag.* **30**, 521–605 (1961).
3. King, G. C., Stein, R. S. & Lin, J. Static stress changes and the triggering of earthquakes. *Bull. Seismol. Soc. Am.* **84**, 935–953 (1994).
4. Toda, S., Stein, R. S., Reasenberg, P. A., Dieterich, J. H. & Yoshida, A. Stress transferred by the 1995 $M_w = 6.9$ Kobe, Japan, shock: effect on aftershocks and future earthquake probabilities. *J. Geophys. Res.* **103**, 24543–24565 (1998).

5. Parsons, T., Stein, R. S., Simpson, R. W. & Reasenberg, P. A. Stress sensitivity of fault seismicity: a comparison between limited-offset oblique and major strike-slip faults. *J. Geophys. Res.* **104**, 20183–20202 (1999).
6. Reasenberg, P. A. & Simpson, R. W. Response of regional seismicity to the static stress change produced by the Loma Prieta earthquake. *Science* **255**, 1687–1690 (1992).
7. Jacques, E., King, G. C. P., Tapponnier, P., Ruegg, J. C. & Manighetti, I. Seismic activity triggered by stress changes after the 1978 events in the Asal Rift, Djibouti. *Geophys. Res. Lett.* **23**, 2481–2484 (1996).
8. Nostro, C., Cocco, M. & Belardinelli, M. E. Static stress changes in extensional regimes: an application to southern Apennines (Italy). *Bull. Seismol. Soc. Am.* **87**, 234–248 (1997).
9. Hardebeck, J. L., Nazareth, J. J. & Hauksson, E. The static stress change triggering model: constraints from two southern California aftershock sequences. *J. Geophys. Res.* **103**, 24427–24437 (1998).
10. Mallman, E. P. & Zoback, M. D. Assessing elastic Coulomb stress transfer models using seismicity rates in southern California and southwestern Japan. *J. Geophys. Res.* **112**, B03304 (2007).
11. Felzer, K. R. & Brodsky, E. E. Testing the stress shadow hypothesis. *J. Geophys. Res.* **110**, B05S09 (2005).
12. Okada, Y. Internal deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **82**, 1018–1040 (1992).
13. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
14. Das, S. & Scholz, C. H. Off-fault aftershock clusters caused by shear stress increase? *Bull. Seismol. Soc. Am.* **71**, 1669–1675 (1981).
15. Kagan, Y. Y. & Jackson, D. D. Spatial aftershock distribution: effect of normal stress. *J. Geophys. Res.* **103**, 24453–24467 (1998).
16. Meade, B. J., DeVries, P., Faller, J., Viegas, F. & Wattenberg, M. What is better than Coulomb failure stress? A ranking of scalar static stress triggering mechanisms from 105 mainshock–aftershock pairs. *Geophys. Res. Lett.* **44**, 11409–11416 (2017).
17. Ma, K. F., Song, T. R. A., Lee, S. J. & Wu, H. I. Spatial slip distribution of the September 20, 1999, Chi-Chi, Taiwan, earthquake ($M_w 7.6$)—inverted from teleseismic data. *Geophys. Res. Lett.* **27**, 3417–3420 (2000).
18. Yoshida, S. et al. Joint inversion of near- and far-field waveforms and geodetic data for the rupture process of the 1995 Kobe earthquake. *J. Phys. Earth* **44**, 437–454 (1996).
19. Shao, G. & Ji, C. Preliminary result of the Oct 8, 2005 Mw 7.6 Pakistan earthquake. UCSB http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2005/10/smooth/2005pakistan.html (accessed 2 June 2018).

Acknowledgements This work was supported by Harvard University and Google. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Reviewer information Nature thanks D. Trugman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors conceived the idea for this paper; P.M.R.D. and B.J.M. implemented the analysis and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0438-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0438-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.M.R.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Neural-network structure and development. To build and train the neural networks, we use the Python toolkit Keras (<https://keras.io>), which provides a high-level application programming interface to access the Theano²⁰ (<http://deeplearning.net/software/theano>) and TensorFlow²¹ (<https://www.tensorflow.org>) deep-learning libraries. We train the networks using Theano, an adaptive learning rate (Adadelta) optimization method²², and a binary cross-entropy cost function. Negative training data (grid cells without aftershocks) are downsampled during training. During training, 10% of the positive training data samples were used for validation, along with an equal number of randomly selected negative samples.

Aftershock catalogue. The aftershocks that occurred between one second and one year after the mainshocks in each grid cell were compiled from the reviewed ISC event catalogue. Note that for seven earthquakes in 2012, the time period included is shorter (as short as about one month for the $M_w = 7.7$ Masset, Canada, earthquake) because the catalogue ends on 30 November 2012.

ROC curve analysis. ROC curves are widely used to evaluate the efficacy of binary classifiers and diagnostic tests in medicine, machine learning and many other fields. To construct an ROC curve, the true positive rate (the ratio of the number of true positive classifications to the total number of positives, or in this case the ratio of the number of grid cells correctly identified as containing aftershocks to the total number of grid cells containing aftershocks) is plotted against the false positive rate (the ratio of the number of false positive classifications to the total number of negatives, or in this case the ratio of the number of grid cells incorrectly identified as containing aftershocks to the number of grid cells that do not contain aftershocks) for the range of possible test thresholds. In this way, an ROC curve represents the performance of a test across all possible thresholds. A binary classification method that is no better than random assignment would plot near the 1:1 line, whereas a test that is more effective than random assignment would plot above the 1:1 line. The AUC quantifies the overall performance of a test; a test that is no better than random assignment would correspond to $AUC = 0.5$, whereas more accurate tests would have AUC values approaching 1. Note that the 2010 Darfield, New Zealand, slip distribution²³ in the test dataset is excluded from the merged AUC value for Coulomb failure stress change (Fig. 2a); the geometric complexity of this rupture precludes meaningful definitions of average strike, rake and dip.

Permutation tests. We perform permutation tests to assess the statistical significance of the performance of the neural-network forecast on the test dataset. For each mainshock slip distribution in the test set, we generated 5,000 random realizations of the locations of positive grid cells. Each of the resulting random ROC curve realizations yields an AUC value. The observed AUC value is compared with the empirical distribution of random AUC values from the permutation tests to obtain a one-sided empirical P value for every slip distribution in the test set.

Quantitative comparison to existing stress metrics for an idealized case. The 40 stress metrics that we consider include the nearest distance to the mainshock rupture, the maximum change in shear stress, the three invariants of the stress-change tensor, the Coulomb failure stress change on receiver planes parallel to the mainshock fault plane (with coefficient of friction $\mu = 0.0, 0.2, 0.4, 0.6$ or 0.8), the total Coulomb failure stress change, the total shear stress change on fault-parallel receiver planes, and the normal-only component of Coulomb failure stress change (see Extended Data Table 1 for mathematical definitions of these quantities). Note that in Figs. 1 and 2a, owing to possible sign ambiguities, we calculate four versions of classic Coulomb failure stress change and use the best-performing sign convention for each slip distribution. In Extended Data Table 1, all of the metrics and their magnitudes are considered for both the deviatoric and full stress tensors in a $300 \text{ km} \times 300 \text{ km}$ area centred on the idealized fault. To enable a quantitative and meaningful comparison of the spatial patterns of these stress metrics to the deep-learning forecast, we first scale and shift each stress metric, then feed these values through a sigmoid filter to normalize between 0 and 1.

Constraining grid-cell depths to 5 km beyond the maximum depth of each slip distribution. The results presented in the main text are based on training and test datasets (Supplementary Table 1; earthquake source data taken from refs 17–19,23–189) that incorporate grid cells down to 50 km for each slip distribution. We originally chose this fixed depth cut-off to keep the analysis as clean, clear and consistent as possible.

In addition, we consider an alternative approach, in which the maximum depth of grid cells considered varies with the maximum depth of each slip distribution. Here we present analogous results to those in the main text, but using training and test datasets that incorporate only grid cells shallower than 5 km below the deepest depth of the slip distribution, although no deeper than 50 km. As in the original training and test datasets, we define these alternative datasets by the random assignments listed in Supplementary Table 1. In other words, for a slip distribution that extends to a depth of 17.5 km, we incorporate all grid cells with centroids at depths shallower than 22.5 km in the test and training datasets specified in Supplementary Table 1.

The size of the training and test datasets are greatly reduced when variable depth limits are incorporated: for the training dataset, the number of total grid cells is

reduced by roughly 20% (from 4,743,090 to 3,779,070), while the total number of aftershocks is also reduced by around 3% (from 131,804 to 127,735). For some individual slip distributions, such as for the Landers earthquake, the number of grid cells is reduced by about 50–60% because the slip distributions extend to depths of 15–18 km, whereas for others—such as for the Tohoku and 2004 Sumatra earthquakes—the number of grid cells incorporated changes by 10% or less because these slip distributions extend to depths of greater than 40 km. We exclude two slip distributions from the analysis when variable depth limits are incorporated: the 2 December 1996 $M_w = 6.7$ Hyuga-nada earthquake (see Supplementary Table 1) in the test dataset and the 1968 $M_w = 7.5$ Hyuga-nada earthquake (see Supplementary Table 1) in the training dataset. When variable depth limits are incorporated for these events, the ISC catalogue does not contain any aftershocks in the volume surrounding the faults.

With these variable-depth-limit training and test datasets, and using a neural network with the same structure as before, we obtain markedly similar results (Extended Data Figs. 2, 3) to those obtained using the original versions of the datasets (Figs. 1, 2). A neural network trained on the variable-depth-limit version of the training dataset yields a merged AUC value across all slip distributions and grid cells in the variable-depth-limit version of the test dataset of 0.8333, which is larger than the classic Coulomb failure stress-change criterion ($AUC = 0.5804$). For this variable-depth-limit case, the neural-network classifier performs similarly well as, although not better than, the maximum change in shear stress ($AUC = 0.8383$) and the von Mises yield criterion ($AUC = 0.8378$). As in the fixed-depth-limit example, more than 98% of the trained-neural-network forward prediction for an idealized strike-slip fault (Extended Data Fig. 3h) can be explained by the maximum change in shear stress, the von Mises yield criterion and the sum of the absolute values of the independent components of the stress-change tensor.

Realizations of the test and training datasets with one slip distribution per mainshock. In the main text, we present results using training and test datasets that are defined by distinct mainshocks, rather than distinct slip distributions. In other words, in each dataset, there may be multiple slip distributions from the same event (see Supplementary Table 1). We took this approach to include as much data as possible; however, certain effects—and perhaps biases—could have been introduced as a result.

To address this concern, here we include results from an ensemble approach to training and testing. In this approach, we generate ten realizations of pairs of training and test datasets. Each realization of the training and test datasets includes only one randomly selected slip distribution per mainshock (Supplementary Table 2). For each of the ten realizations, we used this ensemble approach for both an analysis that incorporated grid cells down to 50 km for each slip distribution and an analysis that incorporated grid cells down to only 5 km beyond the maximum depth of each slip distribution. Thus, a total of 20 networks were trained and tested using the ensemble approach (Supplementary Table 2), with the structure of each the same as that of the original neural network.

Taking this approach substantially reduces the size of the training and test datasets. With grid cells down to 50 km incorporated, the mean number of positive grid cells included in each realization of the training dataset is 33,895.7 (with a minimum number of positive grid cells of 33,670 in realization 7 and a maximum of 34,197 in realization 1). With only grid cells down to 5 km beyond the deepest depth of the slip distributions incorporated, the mean number of positive grid cells included in each realization of the training dataset is further reduced to 31,894.9 (with a minimum number of positive grid cells of 31,162 in realization 5 and a maximum of 32,318 in realization 3). By comparison, in the original training dataset (which incorporates grid cells down to 50 km uniformly), the total number of positive grid cells is 85,850.

Rather than including the ROC curves from all of the realizations of training and test dataset pairs, we instead summarize the relative performances of $\Delta CFS(\mu = 0.4)$, $\Delta \tau_{\max} \sqrt{3\Delta J_2}$ and the associated trained neural networks for each realization (Extended Data Table 2). ROC curves for realization 6—for which the trained neural network performs the worst in terms of the AUC—are shown in Extended Data Fig. 4.

Overall, we obtain similar results to those presented in the main text when using realizations of training and test datasets with one randomly selected slip distribution per event. The AUC values listed in Extended Data Table 2 are comparable to, and in 18 of 20 cases larger than, the AUC values associated with $\Delta \tau_{\max}$ and $\sqrt{3\Delta J_2}$ evaluated on the same test dataset realizations. The smallest AUC value associated with the trained neural networks is 0.78 (realization 6, incorporating only grid cells down to 5 km beyond the deepest depth of each slip distribution; Extended Data Fig. 4) and the largest AUC value associated with $\Delta CFS(\mu = 0.4)$ is 0.626 (realization 6, but incorporating grid cells down to 50 km; Extended Data Fig. 4). Furthermore, the forward predictions of all trained neural networks (Extended Data Figs. 5, 6) are qualitatively similar to those displayed in Fig. 2h and Extended Data Fig. 3h.

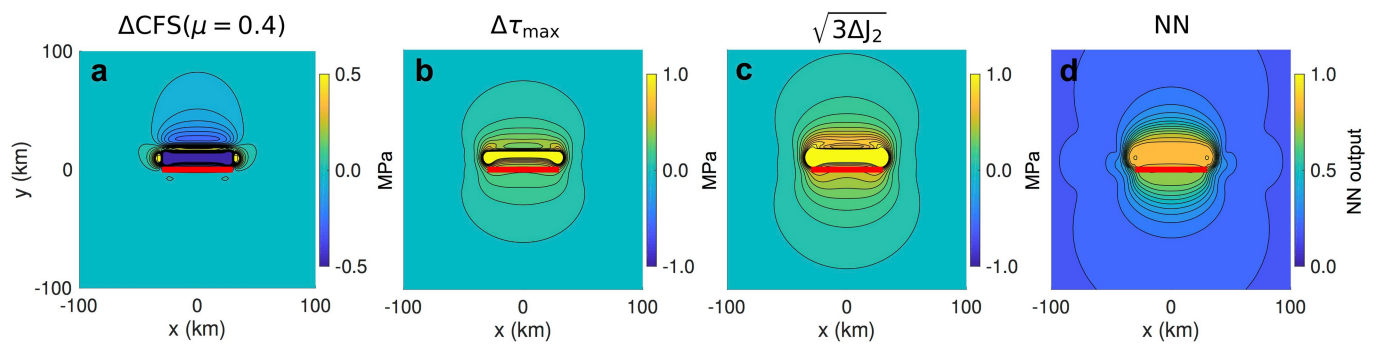
Data and code availability. This project is based on the freely available Keras (<https://keras.io>) and Theano²¹ (<http://deeplearning.net/software/theano>) libraries

- as well as T. B. Thompson's Okada wrapper (https://github.com/tbenthompson/okada_wrapper). All code is available at <https://www.github.com/phoebemrdevries>. All data are freely available from the SRCMOD catalogue (<http://equake-rc.info/SRCMOD>) and the ISC event catalogue (<http://www.isc.ac.uk/iscgem>).
20. The Theano Development Team. Theano: a python framework for fast computation of mathematical expressions. Preprint at <http://arxiv.org/abs/1605.02688> (2016).
 21. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation* 265–283 (USENIX Association, 2016).
 22. Zeiler, M. D. ADDELTA: an adaptive learning rate method. Preprint at <https://arxiv.org/abs/1212.5701> (2012).
 23. Atzori, S. et al. The 2010–2011 Canterbury, New Zealand, seismic sequence: multiple source analysis from InSAR data and modeling. *J. Geophys. Res.* **117**, B08305 (2012).
 24. Yagi, Y., Kikuchi, M., Yoshida, S. & Yamanaka, Y. Source process of the Hyuga-nada Earthquake of April 1, 1968 ($M_{\text{JMA}} 7.5$), and its relationship to the subsequent seismicity. *Zishin J. Seis. Soc. Japan.* **51**, 139–148 (1998) [in Japanese].
 25. Nagai, R., Kikuchi, M. & Yamanaka, Y. Comparative study on the source processes of recurrent large earthquakes in Sariku-oki Region: the 1968 Tokachi-oki earthquake and the 1994 Sanriku-oki earthquake. *Zishin J. Seis. Soc. Japan* **54**, 267–280 (2001) [in Japanese].
 26. Takeo, M. Fault heterogeneity of inland earthquakes in Japan. *Bull. Earthq. Res. Inst. Univ. Tokyo* **65**, 541–569 (1990).
 27. Heaton, T. H. The 1971 San Fernando earthquake: a double event? *Bull. Seismol. Soc. Am.* **72**, 2037–2062 (1982).
 28. Hartzell, S. & Langer, C. Importance of model parameterization in finite fault inversions: application to the 1974 M_w 8.0 Peru earthquake. *J. Geophys. Res.* **98**, 22123–22134 (1993).
 29. Liu, H. & Helmberger, D. V. The near-source ground motion of the 6 August 1979 Coyote Lake, California, earthquake. *Bull. Seismol. Soc. Am.* **73**, 201–218 (1983).
 30. Mendoza, C. Finite-fault analysis of the 1979 March 14 Petatlan, Mexico, earthquake using teleseismic P-wave-forms. *Geophys. J. Int.* **121**, 675–683 (1995).
 31. Takeo, M. Rupture process of the 1980 Izu-Hanto-Toho-Oki earthquake deduced from strong motion seismograms. *Bull. Seismol. Soc. Am.* **78**, 1074–1091 (1988).
 32. Hartzell, S., Langer, C. & Mendoza, C. Rupture histories of eastern North American earthquakes. *Bull. Seismol. Soc. Am.* **84**, 1703–1724 (1994).
 33. Mendoza, C. & Hartzell, S. H. Inversion for slip distribution using teleseismic P waveforms: North Palm Springs, Borah Peak, and Michoacan earthquakes. *Bull. Seismol. Soc. Am.* **78**, 1092–1111 (1988).
 34. Fukuyama, E. & Irikura, K. Rupture process of the 1983 Japan Sea (Akita-Oki) earthquake using a waveform inversion method. *Bull. Seismol. Soc. Am.* **76**, 1623–1640 (1986).
 35. Hartzell, S. H. & Heaton, T. H. Rupture history of the 1984 Morgan Hill, California, earthquake from the inversion of strong motion records. *Bull. Seismol. Soc. Am.* **76**, 649–674 (1986).
 36. Takeo, M. & Mikami, N. Inversion of strong motion seismograms for the source process of the Naganoken-Seibu earthquake of 1984. *Tectonophysics* **144**, 271–285 (1987).
 37. Mendoza, C., Hartzell, S. & Monfret, T. Wide-band analysis of the 3 March 1985 Central Chile earthquake: overall source process and rupture history. *Bull. Seismol. Soc. Am.* **84**, 269–283 (1994).
 38. Mendoza, C. Coseismic slip of two large Mexican earthquakes from teleseismic body wave-forms: implications for asperity interaction in the Michoacan Plate Boundary Segment. *J. Geophys. Res.* **98**, 8197–8210 (1993).
 39. Hartzell, S. Comparison of seismic waveform inversion results for the rupture history of a finite fault: application to the 1986 North Palm-Springs, California, earthquake. *J. Geophys. Res.* **94**, 7515–7534 (1989).
 40. Larsen, S., Reilinger, R., Neugebauer, H. & Strange, W. Global positioning system measurements of deformations associated with the 1987 Superstition Hills earthquake: evidence for conjugate faulting. *J. Geophys. Res.* **97**, 4885–4902 (1992).
 41. Wald, D. J., Helmberger, D. V. & Hartzell, S. H. Rupture process of the 1987 Superstition Hills earthquake from the inversion of strong-motion data. *Bull. Seismol. Soc. Am.* **80**, 1079–1098 (1990).
 42. Hartzell, S. H. & Iida, M. Source complexity of the 1987 Whittier Narrows, California, earthquake from the inversion of strong motion records. *J. Geophys. Res.* **95**, 12475–12485 (1990).
 43. Emolo, A. & Zollo, A. Kinematic source parameters for the 1989 Loma Prieta earthquake from the nonlinear inversion of accelerograms. *Bull. Seismol. Soc. Am.* **95**, 981–994 (2005).
 44. Steidl, J. H., Archuleta, R. J. & Hartzell, S. H. Rupture history of the 1989 Loma Prieta, California, earthquake. *Bull. Seismol. Soc. Am.* **81**, 1573–1602 (1991).
 45. Wald, D. J., Helmberger, D. V. & Heaton, T. H. Rupture model of the 1989 Loma Prieta earthquake from the inversion of strong-motion and broad-band teleseismic data. *Bull. Seismol. Soc. Am.* **81**, 1540–1572 (1991).
 46. Hough, S. E. & Dreger, D. S. Source parameters of the 23 April 1992 M 6.1 Joshua Tree, California, earthquake and its aftershocks: empirical Green's function analysis of GEOS and TERRASCOPE data. *Bull. Seismol. Soc. Am.* **85**, 1576–1590 (1995).
 47. Cohee, B. P. & Beroza, G. C. Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanics. *Bull. Seismol. Soc. Am.* **84**, 692–712 (1994).
 48. Cotton, F. & Campillo, M. Frequency-domain inversion of strong motions: application to the 1992 Landers Earthquake. *J. Geophys. Res.* **100**, 3961–3975 (1995).
 49. Hernandez, B., Cotton, F. & Campillo, M. Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake. *J. Geophys. Res.* **104**, 13083–13099 (1999).
 50. Wald, D. J. & Heaton, T. H. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull. Seismol. Soc. Am.* **84**, 668–691 (1994).
 51. Zeng, Y. & Anderson, J. *Evaluation of Numerical Procedures for Simulating Near-Fault Long-Period Ground Motions using Zeng Method*. Report No. 2000/01 to the PEER Utilities Program (Pacific Earthquake Engineering Research Center, UC Berkeley, 2000).
 52. Mendoza, C. & Fukuyama, E. The July 12, 1993, Hokkaido-Nansei-Oki, Japan, earthquake: coseismic slip pattern from strong-motion and teleseismic recordings. *J. Geophys. Res.* **101**, 791–801 (1996).
 53. Hartzell, S., Liu, P. C. & Mendoza, C. The 1994 Northridge, California, earthquake; investigation of rupture velocity, risetime, and high-frequency radiation. *J. Geophys. Res.* **101**, 20091–20108 (1996).
 54. Hudnut, K. W. et al. Co-seismic displacements of the 1994 Northridge, California, earthquake. *Bull. Seismol. Soc. Am.* **86**, S19–S36 (1996).
 55. Shen, Z.-K. et al. Northridge earthquake rupture models based on the global positioning system measurements. *Bull. Seismol. Soc. Am.* **86**, S37–S48 (1996).
 56. Wald, D. J., Heaton, T. H. & Hudnut, K. W. The slip history of the 1994 Northridge, California, earthquake determined from strong-motion, teleseismic, GPS, and leveling data. *Bull. Seismol. Soc. Am.* **86**, S49–S70 (1996).
 57. Nakayama, W. & Takeo, M. Slip history of the 1994 Sanriku-Haruka-Oki, Japan, earthquake deduced from strong-motion data. *Bull. Seismol. Soc. Am.* **87**, 918–931 (1997).
 58. Mendoza, C. & Hartzell, S. Fault-slip distribution of the 1995 Colima-Jalisco, Mexico, earthquake. *Bull. Seismol. Soc. Am.* **89**, 1338–1344 (1999).
 59. Courbouloux, F., Santoyo, M. A., Pacheco, J. F. & Singh, S. K. The 14 September 1995 ($M = 7.3$) Copala, Mexico, earthquake: a source study using teleseismic, regional, and local data. *Bull. Seismol. Soc. Am.* **87**, 999–1010 (1997).
 60. Yagi, Y., Kikuchi, M., Yoshida, S. & Sagiya, T. Comparison of the coseismic rupture with the aftershock distribution in the Hyuga-nada earthquakes of 1996. *Geophys. Res. Lett.* **26**, 3161–3164 (1999).
 61. Salichon, J. et al. Joint inversion of broadband teleseismic and interferometric synthetic aperture radar (InSAR) data for the slip history of the $M_w = 7.7$, Nazca ridge (Peru) earthquake of 12 November 1996. *J. Geophys. Res.* **108**, 2085 (2003).
 62. Hernandez, B. et al. Rupture history of the 1997 Umbria-Marche (central Italy) main shocks from the inversion of GPS, DInSAR and near field strong motion data. *Ann. Geophys.* **47**, 1355–1376 (2004).
 63. Horikawa, H. Earthquake doublet in Kagoshima, Japan: rupture of asperities in a stress shadow. *Bull. Seismol. Soc. Am.* **91**, 112–127 (2001).
 64. Sudhaus, H. & Jönsson, S. Source model for the 1997 Zirkuh earthquake ($M_w = 7.2$) in Iran derived from JERS and ERS InSAR observations. *Geophys. J. Int.* **185**, 676–692 (2011).
 65. Ide, S. Complex source processes and the interaction of moderate earthquakes during the earthquake swarm in the Hida-Mountains, Japan, 1998. *Tectonophysics* **334**, 35–54 (2001).
 66. Miyakoshi, K., Kagawa, T., Sekiguchi, H., Iwata, T. & Irikura, K. Source characterization of inland earthquakes in Japan using source inversion results. In *Proc. 12th World Conference on Earthquake Engineering abstr.* 1850 (New Zealand Society for Earthquake Engineering, 2000).
 67. Nakahara, H. et al. Broadband source process of the 1998 Iwate prefecture, Japan, earthquake as revealed from inversion analyses of seismic waveforms and envelopes. *Bull. Seismol. Soc. Am.* **92**, 1708–1720 (2002).
 68. Ji, C., Wald, D. J. & Helmberger, D. V. Source description of the 1999 Hector Mine, California, earthquake, part II: complexity of slip history. *Bull. Seismol. Soc. Am.* **92**, 1208–1226 (2002).
 69. Salichon, J., Lundgren, P., Delouis, B. & Giardini, D. Slip history of the 16 October 1999 M_w 7.1 Hector Mine earthquake (California) from the inversion of InSAR, GPS, and teleseismic data. *Bull. Seismol. Soc. Am.* **94**, 2015–2027 (2004).
 70. Bouchon, M. et al. Space and time evolution of rupture and faulting during the 1999 Izmit (Turkey) earthquake. *Bull. Seismol. Soc. Am.* **92**, 256–266 (2002).
 71. Çakir, Z. et al. Coseismic and early post-seismic slip associated with the 1999 İzmit earthquake (Turkey), from SAR interferometry and tectonic field observations. *Geophys. J. Int.* **155**, 93–110 (2003).
 72. Delouis, B., Giardini, D., Lundgren, P. & Salichon, J. Joint inversion of InSAR, GPS, teleseismic, and strong-motion data for the spatial and temporal distribution of earthquake slip: application to the 1999 İzmit mainshock. *Bull. Seismol. Soc. Am.* **92**, 278–299 (2002).
 73. Reilinger, R. E. et al. Coseismic and postseismic fault slip for the 17 August 1999, $M = 7.5$, İzmit, Turkey earthquake. *Science* **289**, 1519–1524 (2000).
 74. Yagi, Y. & Kikuchi, M. Source rupture process of the Kocaeli, Turkey, earthquake of August 17, 1999, obtained by joint inversion of near-field data and teleseismic data. *Geophys. Res. Lett.* **27**, 1969–1972 (2000).
 75. Copley, A., Avouac, J. P., Hollingsworth, J. & Leprince, S. The 2001 M_w 7.6 Bhuj earthquake, low fault friction, and the crustal support of plate driving forces in India. *J. Geophys. Res.* **116**, B08405 (2011).

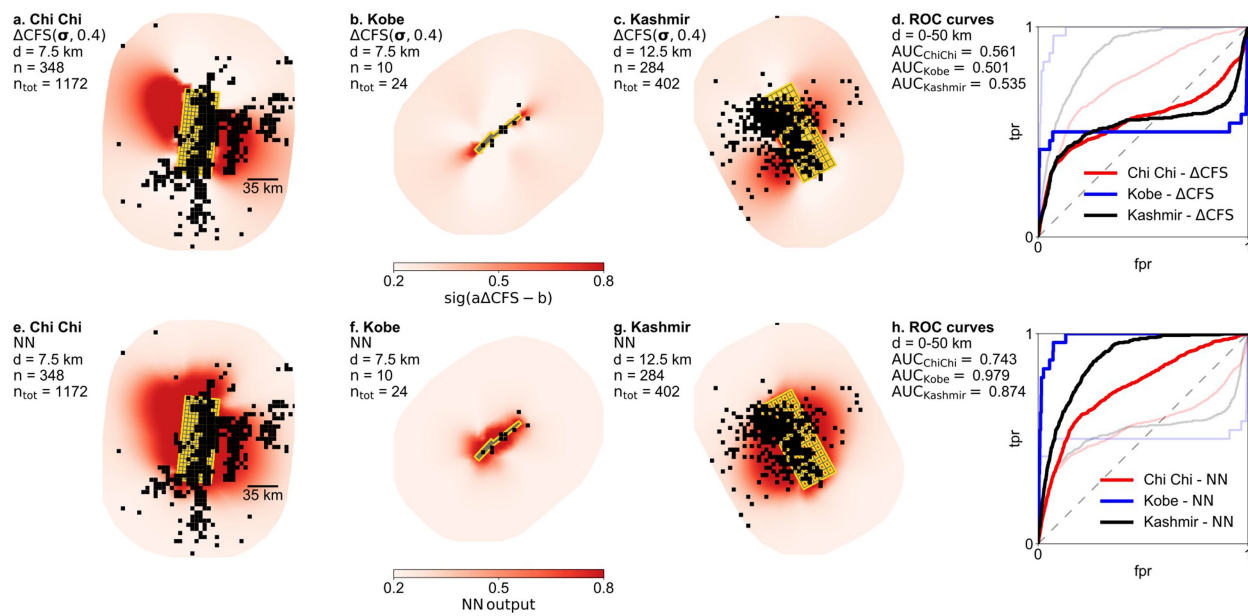
76. Copley, A. Source models of large earthquakes: Jan/26/2001 (Mw 7.6), Bhuj, India. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2001_Bhuj/index.html (accessed 1 July 2013).
77. Yagi, Y. A slip model for the Jan 26, 2001 Bhuj (India) earthquake using teleseismic recordings. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2001BHUIJN01YAGI> (accessed 18 May 2003).
78. Asano, K., Iwata, T. & Irikura, K. Estimation of source rupture process and strong ground motion simulation of the 2002 Denali, Alaska, earthquake. *Bull. Seismol. Soc. Am.* **95**, 1701–1715 (2005).
79. Poiata, N., Miyake, H., Koketsu, K. & Hikima, K. Strong motion and teleseismic waveform inversions for the source process of the 2003 Bam, Iran, earthquake. *Bull. Seismol. Soc. Am.* **102**, 1477–1496 (2012).
80. Semmane, F., Campillo, M. & Cotton, F. Fault location and source process of the Boumerdes, Algeria, earthquake inferred from geodetic and strong motion data. *Geophys. Res. Lett.* **32**, L01305 (2005).
81. Wei, S. Source models of large earthquakes: July/15/2003, Carlsberg Ridge, Mw7.6. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2004_carlsberg-ridge/index.html (accessed 1 July 2013).
82. Koketsu, K., Hikima, K., Miyazaki, S. & Ide, S. Joint inversion of strong motion and geodetic data for the source process of the 2003 Tokachi-oki, Hokkaido, earthquake. *Earth Planets Space* **56**, 329–334 (2004).
83. Tanioka, Y., Hirata, K., Hino, R. & Kanazawa, T. Slip distribution of the 2003 Tokachi-oki earthquake estimated from tsunami waveform inversion. *Earth Planets Space* **56**, 373–376 (2004).
84. Yagi, Y. Source rupture process of the 2003 Tokachi-oki earthquake determined by joint inversion of teleseismic body wave and strong ground motion data. *Earth Planets Space* **56**, 311–316 (2004).
85. Yamanaka, Y. & Kikuchi, M. Source process of the recurrent Tokachi-oki earthquake on September 26, 2003, inferred from teleseismic body waves. *Earth Planets Space* **55**, e21–e24 (2003).
86. Wei, S. Source models of large earthquakes: Feb/07/2004 (Mw 7.2), Irian Jaya, Indonesia. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2004_indo-irian_jaya/index.html (accessed 1 July 2013).
87. Ammon, C. J. et al. Rupture process of the great 2004 Sumatra-Andaman earthquake. *Science* **308**, 1133–1139 (2005).
88. Rhie, J., Dreger, D., Burgmann, R. & Romanowicz, B. Slip of the 2004 Sumatra-Andaman earthquake from joint inversion of long-period global seismic waveforms and GPS static offsets. *Bull. Seismol. Soc. Am.* **97**, S115–S127 (2007).
89. Shao, G. & Ji, C. Preliminary result of the Aug 16, 2005 Mw 7.19 Honshu earthquake. *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2005/08/smooth/honshu.html (accessed 22 August 2013).
90. Shao, G. & Ji, C. Preliminary result of the Jun 15, 2005 Mw 7.2 northern California earthquake. *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2005/06/smooth/northernca.html (accessed 25 September 2013).
91. Lay, T. et al. The 2006–2007 Kuril Islands great earthquake sequence. *J. Geophys. Res.* **114**, B11308 (2009).
92. Sladen, A. Source models of large earthquakes: preliminary result, 11/15/2006 (Mw 8.3), Kuril Islands. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2006_kuril/kuril.html (accessed 1 July 2013).
93. Yen, Y.-T., Ma, K.-F. & Wen, Y.-Y. Slip partition of the 26 December 2006 Pingtung, Taiwan (M6.9, M6.8) earthquake doublet determined from teleseismic waveforms. *Diqiu Kexue Jikan* **19**, 567–578 (2008).
94. Ji, C. Rupture process of the 2007 Jan 13 magnitude 8.1 - KURIL Island earthquake (revised). *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2007/01/13/kuril.html (accessed 22 August 2013).
95. Sladen, A. Source models of large earthquakes: preliminary result, 01/13/2007 (Mw 8.1), Kuril Islands. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2007_kuril/kuril.html (accessed 1 July 2013).
96. Ji, C. & Zeng, Y. Preliminary result of the Sep 12, 2007 Mw 7.9 Kepulauan earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2007PAGAI01Jlxx> (accessed 14 June 2018).
97. Sladen, A. & Ozgun Konca, A. Source models of large earthquakes: preliminary result, 09/12/2007 (Mw 7.9), Central Sumatra earthquake. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2007_c-sumatra/c-sumatra.html (accessed 1 July 2013).
98. Ji, C. Rupture process of the 2007 April 1, magnitude 8.1, Solomon Islands earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2007SOLOMO01Jlxx> (accessed 14 June 2018).
99. Béjar-Pizarro, M. et al. Asperities and barriers on the seismogenic zone in North Chile: state-of-the-art after the 2007 M_w 7.7 Tocopilla earthquake inferred by GPS and InSAR data. *Geophys. J. Int.* **183**, 390–406 (2010).
100. Sladen, A. Source models of large earthquakes: preliminary result, 11/14/2007 (Mw 7.7), Tocopilla earthquake, Chile. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2007_tocopilla/tocopilla.html (accessed 1 July 2013).
101. Zeng, Y., Hayes, G. & Ji, C. Preliminary result of the Nov 14, 2007 Mw 7.7 Antofagasto, Chile earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2007TOCOP101ZENG> (accessed 14 June 2018).
102. Sladen, A. Source models of large earthquakes: preliminary result, 11/16/2008 (Mw 7.3), Sulawesi. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2008_sulawesi/sulawesi.html (accessed 1 July 2013).
103. Ji, C. & Hayes, G. Preliminary result of the May 12, 2008 Mw 7.9 eastern Sichuan, China earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2008WENCHU01Jlxx> (accessed 18 June 2018).
104. Sladen, A. Source models of large earthquakes: preliminary result, 05/12/2008 (Mw 7.9), East Sichuan. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2008_e_sichuan/e_sichuan.html (accessed 1 July 2013).
105. Yagi, Y., Nishimura, N. & Kasahara, A. Source process of the 12 May 2008 Wenchuan, China, earthquake determined by waveform inversion of teleseismic body waves with a data covariance matrix. *Earth Planets Space* **64**, e13–e16 (2012).
106. Fielding, E. J. et al. Kinematic fault slip evolution source models of the 2008 M7.9 Wenchuan earthquake in China from SAR interferometry, GPS and teleseismic analysis and implications for Longmen Shan tectonics. *Geophys. J. Int.* **194**, 1138–1166 (2013).
107. Hayes, G. Preliminary result of the July 15, 2009 Mw 7.6 Fiordland earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009FIORDL01HAYE> (accessed 14 June 2018).
108. Hayes, G. Preliminary result of the August 3, 2009 Mw 6.9 Gulf of California earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009GULFOF01HAYE> (accessed 14 June 2018).
109. Cirella, A., Piatanesi, A., Tinti, E., Chini, M. & Cocco, M. Complexity of the rupture process during the 2009 L'Aquila, Italy, earthquake. *Geophys. J. Int.* **190**, 607–621 (2012).
110. Gualandi, A., Serpelloni, E. & Belardinelli, M. E. s2009LAQUIL01GUAL. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009LAQUIL01GUAL> (accessed June 2018).
111. Hayes, G. & Ji, C. Preliminary result of the May 28, 2009 Mw 7.3 earthquake offshore Honduras. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009OFFSHO01HAYE> (accessed 14 June 2018).
112. Hayes, G. A preliminary result of the Sep 30, 2009 Mw 7.6 southern Sumatra earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009PADANG01HAYE> (accessed 19 June 2018).
113. Sladen, A. Source models of large earthquakes: preliminary result, 09/30/2009 (Mw 7.6), Padang, Indonesia. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2009_padang/padang.html (accessed 1 July 2013).
114. Hayes, G. Preliminary result of the Sep 29, 2009 Mw 8.0 Samoa earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009SAMOA01HAYE> (accessed 19 June 2018).
115. Sladen, A. Source models of large earthquakes: preliminary result, 10/07/2009 (Mw 7.6), Vanuatu. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2009_vanuatu/index.html (accessed 1 July 2013).
116. Wei, S. et al. Superficial simplicity of the 2010 El Mayor–Cuicapah earthquake of Baja California in Mexico. *Nat. Geosci.* **4**, 615–618 (2011).
117. Calais, E. et al. Transpressional rupture of an unmapped fault during the 2010 Haiti earthquake. *Nat. Geosci.* **3**, 794–799 (2010).
118. Hayes, G. P. et al. Complex rupture during the 12 January 2010 Haiti earthquake. *Nat. Geosci.* **3**, 800–805 (2010).
119. Sladen, A. Source models of large earthquakes: preliminary result, 01/12/2010 (Mw 7.0), Haiti. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2010_haiti/index.html (accessed 1 July 2013).
120. Hayes, G. Updated result of the Jan 12, 2010 Mw 7.0 Haiti earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010HAITI02HAYE> (accessed 20 June 2018).
121. Hayes, G. Updated result of the Feb 27, 2010 Mw 8.8 Maule, Chile earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010MAULEC01HAYE> (accessed 19 June 2018).
122. Hayes, G. Updated result of the Apr 6, 2010 northern Sumatra earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010NORTHE01HAYE> (accessed 19 June 2018).
123. Hayes, G. Preliminary result of the Dec 25, 2010 Mw 7.3 Vanuatu region earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010VANUAT01HAYE> (accessed 19 June 2018).
124. Hayes, G. Preliminary result of the Oct 21, 2011 Mw 7.4 Kermadec Islands region earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2011KERMADO2HAYE> (accessed 3 June 2018).
125. Hayes, G. Preliminary result of the July 6, 2011 Mw 7.6 Kermadec Islands region earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2011KERMADO1HAYE> (accessed 1 June 2018).
126. Hayes, G. Updated result of the Mar 9, 2011 Mw 7.3 earthquake offshore Honshu, Japan (Tohoku EQ foreshock). *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2011OFFSHO01HAYE> (accessed 1 June 2018).

127. Fujii, Y., Satake, K., Sakai, S., Shinohara, M. & Kanazawa, T. Tsunami source of the 2011 off the Pacific coast of Tohoku earthquake. *Earth Planets Space* **63**, 815–820 (2011).
128. Satake, K., Fujii, Y., Harada, T. & Namegaya, Y. Time and space distribution of coseismic slip of the 2011 Tohoku earthquake as inferred from tsunami waveform data. *Bull. Seismol. Soc. Am.* **103**, 1473–1492 (2013).
129. Yue, H. & Lay, T. Source rupture models for the M_w 9.0 2011 Tohoku earthquake from joint inversions of high-rate geodetic and seismic data. *Bull. Seismol. Soc. Am.* **103**, 1242–1255 (2013).
130. Hayes, G. Updated result of the Oct 23, 2011 Mw 7.1 eastern Turkey earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2011VANTUR01HAYE> (accessed 20 June 2018).
131. Shao, G. & Ji, C. Preliminary result of the Oct 23, 2011 Mw 7.13 Turkey earthquake. *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2011/10/23/turkey.html (accessed 22 August 2013).
132. Hayes, G. Preliminary result of the Aug 20, 2011 Mw 7.1 Vanuatu earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2011VANUAT01HAYE> (accessed 1 June 2018).
133. Wei, S. et al. Complementary slip distributions of the largest earthquakes in the 2012 Brawley swarm, Imperial Valley, California. *Geophys. Res. Lett.* **40**, 847–852 (2013).
134. Hayes, G. Preliminary result of the Aug 31, 2012 Mw 7.6 earthquake east of Sulang, Philippines. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012EASTOF01HAYE> (accessed 21 June 2018).
135. Lay, T. et al. The October 28, 2012 M_w 7.8 Haida Gwaii underthrusting earthquake and tsunami: slip partitioning along the Queen Charlotte fault transpressional plate boundary. *Earth Planet. Sci. Lett.* **375**, 57–70 (2013).
136. Shao, G. & Ji, C. Preliminary result of the Oct 28, 2012 Mw 7.72 Canada earthquake. *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2012/10/canada.html (accessed 20 August 2013).
137. Wei, S. Source models of large earthquakes: Oct/28/2012 (M_w 7.8), Masset, Canada. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2012_Masset/index.html (accessed 1 July 2013).
138. Hayes, G. Preliminary result of the Mar 20, 2012 Mw 7.4 Oaxaca, Mexico earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012OAXACA01HAYE> (accessed 20 June 2018).
139. Wei, S. Source models of large earthquakes: March/20/2012 (M_w 7.4), OAXACA, Mexico. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2012_Mexico/index.html (accessed 1 July 2013).
140. Hayes, G. Preliminary result of the Jan 10, 2012 Mw 7.2 off the west coast of northern Sumatra, Indonesia earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012SUMATRO3HAYE> (accessed 14 June 2018).
141. Shao, G., Li, X. & Ji, C. Preliminary result of the Apr 11, 2012 Mw 8.64 Sumatra earthquake. *UCSB* http://www.geol.ucsb.edu/faculty/ji/big_earthquakes/2012/04/10/sumatra.html (accessed 19 August 2013).
142. Hayes, G. Preliminary result of the Apr 11, 2012 Mw 8.6 earthquake off the west coast of northern Sumatra. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012SUMATRO1HAYE> (accessed 19 June 2018).
143. Hayes, G. Preliminary result of the Apr 11, 2012 Mw 8.6 earthquake off the west coast of northern Sumatra. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012SUMATRO2HAYE> (accessed 19 June 2018).
144. Yamanaka, Y. & Kikuchi, M. Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. *J. Geophys. Res.* **109**, B07307 (2004).
145. Hartzell, S. & Mendoza, C. Application of an iterative least-squares wave-form inversion of strong-motion and teleseismic records to the 1978 Tabas, Iran, earthquake. *Bull. Seismol. Soc. Am.* **81**, 305–331 (1991).
146. Archuleta, R. J. A faulting model for the 1979 Imperial Valley earthquake. *J. Geophys. Res.* **89**, 4559–4585 (1984).
147. Hartzell, S. H. & Heaton, T. H. Inversion of strong ground motion and teleseismic waveform data for the fault rupture history of the 1979 Imperial Valley, California, earthquake. *Bull. Seismol. Soc. Am.* **73**, 1553–1583 (1983).
148. Olson, A. H. & Apsel, R. J. Finite faults and inverse theory with applications to the 1979 Imperial Valley earthquake. *Bull. Seismol. Soc. Am.* **72**, 1969–2001 (1982).
149. Mendoza, C. & Hartzell, S. H. Slip distribution of the 19 September 1985 Michoacan, Mexico, earthquake: near-source and teleseismic constraints. *Bull. Seismol. Soc. Am.* **79**, 655–669 (1989).
150. Wald, D. J. Strong motion and broad-band teleseismic analysis of the 1991 Sierra-Madre, California, earthquake. *J. Geophys. Res.* **97**, 11033–11046 (1992).
151. Silva, W. et al. A slip model for the Little Skull Mountain earthquake of June 29, 1992. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s1992LITTLE01SILV> (accessed June 2018).
152. Cho, I. & Nakanishi, I. Investigation of the three-dimensional fault geometry ruptured by the 1995 Hyogo-Ken Nanbu earthquake using strong-motion and geodetic data. *Bull. Seismol. Soc. Am.* **90**, 450–467 (2000).
153. Horikawa, H., Hirahara, K., Umeda, Y., Hashimoto, M. & Kusano, F. Simultaneous inversion of geodetic and strong-motion data for the source process of the Hyogo-ken Nanbu, Japan, earthquake. *J. Phys. Earth* **44**, 455–471 (1996).
154. Ide, S., Takeo, M. & Yoshida, Y. Source process of the 1995 Kobe earthquake: determination of spatio-temporal slip distribution by Bayesian modeling. *Bull. Seismol. Soc. Am.* **86**, 547–566 (1996).
155. Koketsu, K., Yoshida, S. & Higashihara, H. A fault model of the 1995 Kobe earthquake derived from the GPS data on the Akashi Kaikyo Bridge and other datasets. *Earth Planets Space* **50**, 803–811 (1998).
156. Sekiguchi, H., Irikura, K. & Iwata, T. Fault geometry at the rupture termination of the 1995 Hyogo-ken Nanbu earthquake. *Bull. Seismol. Soc. Am.* **90**, 117–133 (2000).
157. Wald, D. J. Slip history of the 1995 Kobe, Japan, earthquake determined from strong motion, teleseismic, and geodetic data. *J. Phys. Earth* **44**, 489–503 (1996).
158. Sekiguchi, H., Irikura, K. & Iwata, T. Source inversion for estimating the continuous slip distribution on a fault introduction of Green's functions convolved with a correction function to give moving dislocation effects in subfaults. *Geophys. J. Int.* **150**, 377–391 (2002).
159. Chi, W. C., Dreger, D. & Kaverina, A. Finite-source modeling of the 1999 Taiwan (Chi-Chi) earthquake derived from a dense strong-motion network. *Bull. Seismol. Soc. Am.* **91**, 1144–1157 (2004).
160. Jonsson, S., Zebker, H., Segall, P. & Amelung, F. Fault slip distribution of the 1999 M_w 7.1 Hector Mine, California, earthquake, estimated from satellite radar and GPS measurements. *Bull. Seismol. Soc. Am.* **92**, 1377–1389 (2002).
161. Zhang, W., Iwata, T., Irikura, K., Pitarka, A. & Sekiguchi, H. Dynamic rupture process of the 1999 Chi-Chi, Taiwan, earthquake. *Geophys. Res. Lett.* **31**, L10605 (2004).
162. Wu, C. J., Takeo, M. & Ide, S. Source process of the Chi-Chi earthquake: a joint inversion of strong motion data and global positioning system data with a multifault model. *Bull. Seismol. Soc. Am.* **91**, 1128–1143 (2004).
163. Zeng, Y. H. & Chen, C. H. Fault rupture process of the 20 September 1999 Chi-Chi, Taiwan, earthquake. *Bull. Seismol. Soc. Am.* **91**, 1088–1098 (2004).
164. Ma, K. F., Mori, J., Lee, S. J. & Yu, S. B. Spatial and temporal distribution of slip for the 1999 Chi-Chi, Taiwan, earthquake. *Bull. Seismol. Soc. Am.* **91**, 1069–1087 (2004).
165. Birgören, G., Sekiguchi, H. & Irikura, K. Rupture model of the 1999 Düzce, Turkey, earthquake deduced from high and low frequency strong motion data. *Geophys. Res. Lett.* **31**, L05610 (2004).
166. Delouis, B., Lundgren, P. & Giardini, D. Slip distributions of the 1999 Düzce (M_w 7.2) and Izmit (M_w 7.6) earthquakes on the North Anatolian Fault (Turkey): a combined analysis, internal report. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s1999DUZCETO1DELO> (accessed 1 April 2018).
167. Hernandez, B. et al. Rupture history of September 30, 1999 intraplate earthquake of Oaxaca, Mexico (M_w 7.5) from inversion of strong-motion data. *Geophys. Res. Lett.* **28**, 363–366 (2001).
168. Iwata, T., Sekiguchi, H., Matsumoto, Y., Miyake, H. & Irikura, K. Source process of the 2000 western Tottori Prefecture earthquake and near-source strong ground motion. In *2000 Fall Meeting of the Seismological Society of Japan* (Seismological Society of Japan, 2000).
169. Sekiguchi, H., Iwata, T., Sugiyama, Y., Fusejima, Y. & Horikawa, H. Faulting process and condition for its occurrence of 2000 Tottori-ken Seibu Earthquake. In *2001 Japan Earth and Planetary Science Joint Meeting abstr.* S3-006 (2001).
170. Kakehi, Y. Analysis of the 2001 Geiyo, Japan, earthquake using high-density strong ground motion data: detailed rupture process of a slab earthquake in a medium with a large velocity contrast. *J. Geophys. Res.* **109**, B08306 (2004).
171. Yagi, Y., Mikurno, T., Pacheco, J. & Reyes, G. Source rupture process of the Tecoman, Colima, Mexico earthquake of 22 January 2003, determined by joint inversion of teleseismic body-wave and near-source data. *Bull. Seismol. Soc. Am.* **94**, 1795–1807 (2004).
172. Custódio, S., Liu, P. C. & Archuleta, R. J. The 2004 M_w 6.0 Parkfield, California, earthquake: inversion of near-source ground motion using multiple data sets. *Geophys. Res. Lett.* **32**, L23312 (2005).
173. Dreger, D. S., Gee, L., Lombard, P., Murray, M. H. & Romanowicz, B. Rapid finite-source analysis and near-fault strong ground motions: application to the 2003 M_w 6.5 San Simeon and 2004 M_w 6.0 Parkfield earthquakes. *Seismol. Res. Lett.* **76**, 40–48 (2005).
174. Ji, C. Source models of large earthquakes: slip history the 2004 (M_w 5.9) Parkfield earthquake (single-plane model). *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2004_ca/parkfield2.html (accessed 1 July 2013).
175. Ozgun Konca, A. Source models of large earthquakes: preliminary result, 06/10/08 (M_w 7.6), Kashmir earthquake. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2005_kashmir/kashmir.html (accessed 1 July 2013).
176. Yagi, Y. & Fukahata, Y. Rupture process of the 2011 Tohoku-oki earthquake and absolute elastic strain release. *Geophys. Res. Lett.* **38**, L19307 (2011).
177. Ji, C. Preliminary result of the 2006 July 17 magnitude 7.7 - south of Java, Indonesia earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2006SOUTH01Jlxx> (accessed 21 June 2018).
178. Ozgun Konca, A. Source models of large earthquakes: preliminary result, 06/07/17 (M_w 7.9), southern Java earthquake. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2006_s_java/s_java.html (accessed 1 July 2013).

179. Ji, C. & Zeng, Y. Preliminary result of the Aug 15, 2007 Mw 8.0 coast of central Peru earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2007PISCOP01Jlxx> (accessed 21 June 2018).
180. Ozgun Konca, A. Source models of large earthquakes: preliminary result, 07/08/15 (Mw 8.0), Peru earthquake. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2007_peru/pisco.html (accessed 1 July 2013).
181. Hayes, G. & Ji, C. Preliminary result of the Jun 13, 2008 Mw 6.8 Honshu earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2008IWATEX01HAYE> (accessed 28 June 2018).
182. Hayes, G. Preliminary result of the Sep 29, 2008 Mw 7.0 Kermadec Islands earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2008KERMED01HAYE> (accessed 19 June 2013).
183. Hayes, G. & Ji, C. Preliminary result of the Feb 20, 2008 Mw 7.4 Simeulue earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2008SIMEUL01HAYE> (accessed 20 June 2018).
184. Sladen, A. Source models of large earthquakes: preliminary result 02/20/2008 (Mw 7.4), Simeulue earthquake, Indonesia. *Caltech Tectonics Observatory* http://www.tectonics.caltech.edu/slip_history/2008_n_sumatra/simeulue.html (accessed 1 July 2013).
185. Hayes, G. Preliminary result of the Jan 3, 2009 Mw 7.6 Papua earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2009PAPUAx01HAYE> (accessed 19 June 2018).
186. Hayes, G. Preliminary result of the Dec 21, 2010 Mw 7.4 Bonin Islands earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010BONINI01HAYE> (accessed 19 June 2018).
187. Hayes, G. Preliminary result of the May 9, 2010 Mw 7.2 northern Sumatra earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2010NORTHE02HAYE> (accessed 19 June 2018).
188. Hayes, G. Preliminary result of the Sep 5, 2012 Mw 7.6 Costa Rica earthquake. *eQuake-RC Finite-Source Rupture Model Database* <http://equake-rc.info/SRCMOD/searchmodels/viewmodel/s2012COSTAR01HAYE> (accessed 21 June 2018).
189. Yue, H. et al. The 5 September 2012 Nicoya, Costa Rica M_w 7.6 earthquake rupture process from joint inversion of high-rate GPS, strong-motion, and teleseismic P wave data and its relationship to adjacent plate boundary interface properties. *J. Geophys. Res.* **118**, 5453–5466 (2013).

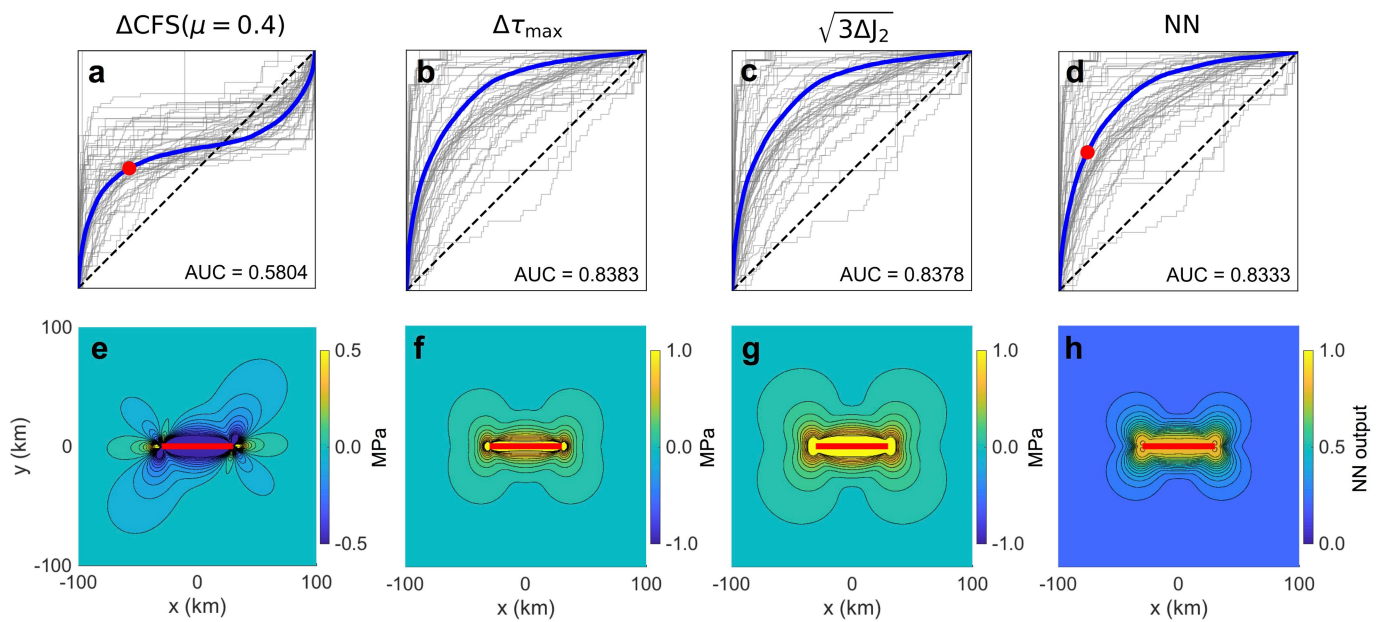


Extended Data Fig. 1 | Comparisons of spatial patterns of stress metrics. a–d, Analogous to Fig. 2e–h, but for an idealized thrust earthquake. The fault plane dips 45° to the north and the red line is the trace of the fault at the surface. Depth shown is 10 km.



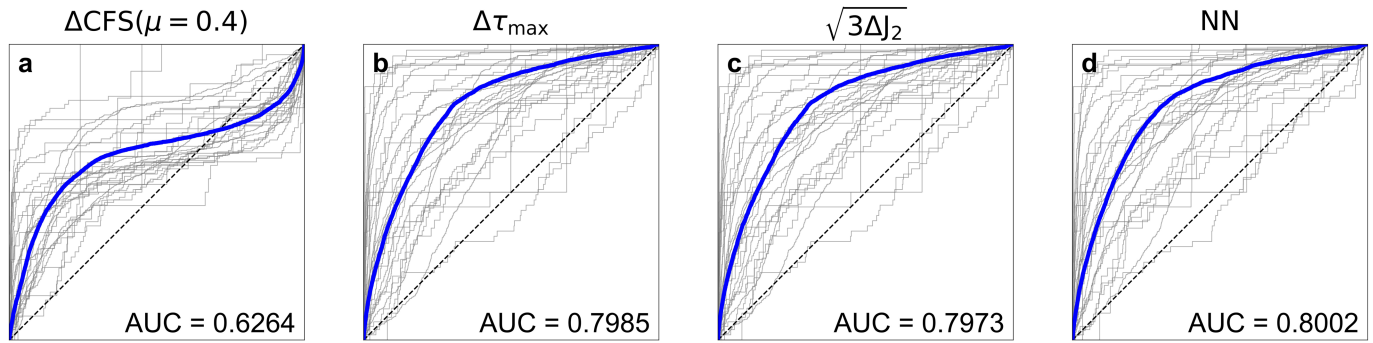
Extended Data Fig. 2 | Mainshock-aftershock examples. a–h, Analogous to Fig. 1a–h, using the same sign conventions for Coulomb failure stress change, but with results based on a training dataset (Supplementary

Table 1) that excludes grid cells more than 5 km below the maximum depth of each slip distribution.

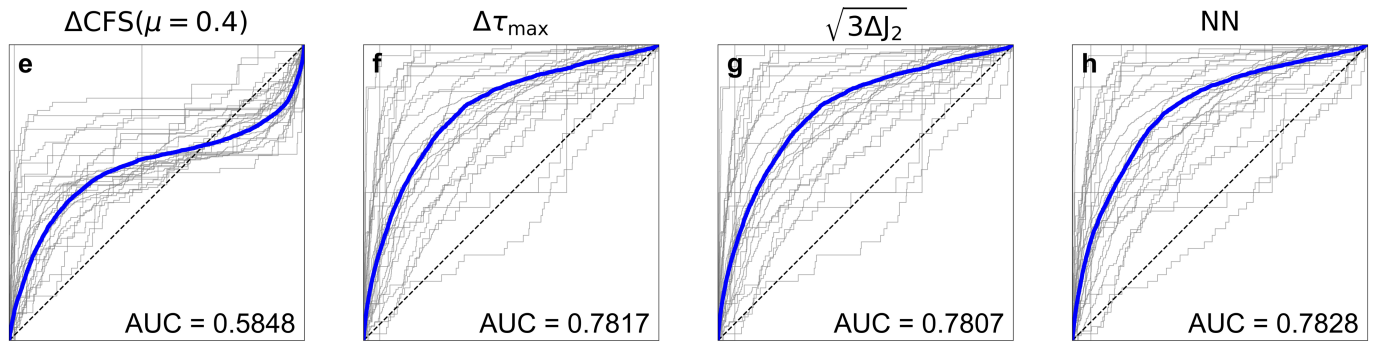


Extended Data Fig. 3 | Comparisons of performance. a–h, Analogous to Fig. 2a–h, using training and test datasets (Supplementary Table 1) that exclude grid cells more than 5 km below the maximum depth of each slip distribution.

Including grid cells to 50 km depth

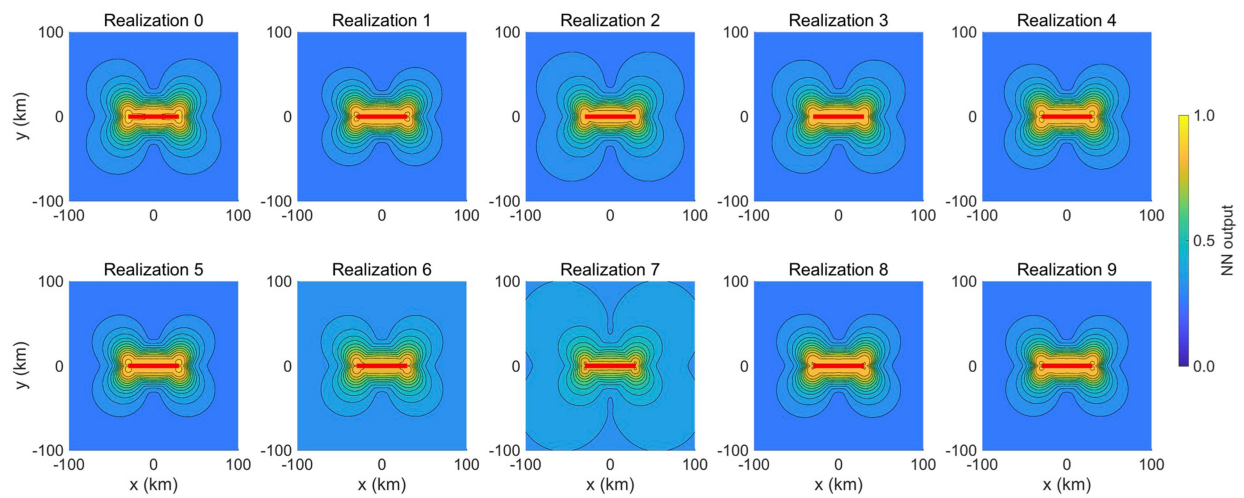


Including grid cells to 5 km beyond depth extent of fault



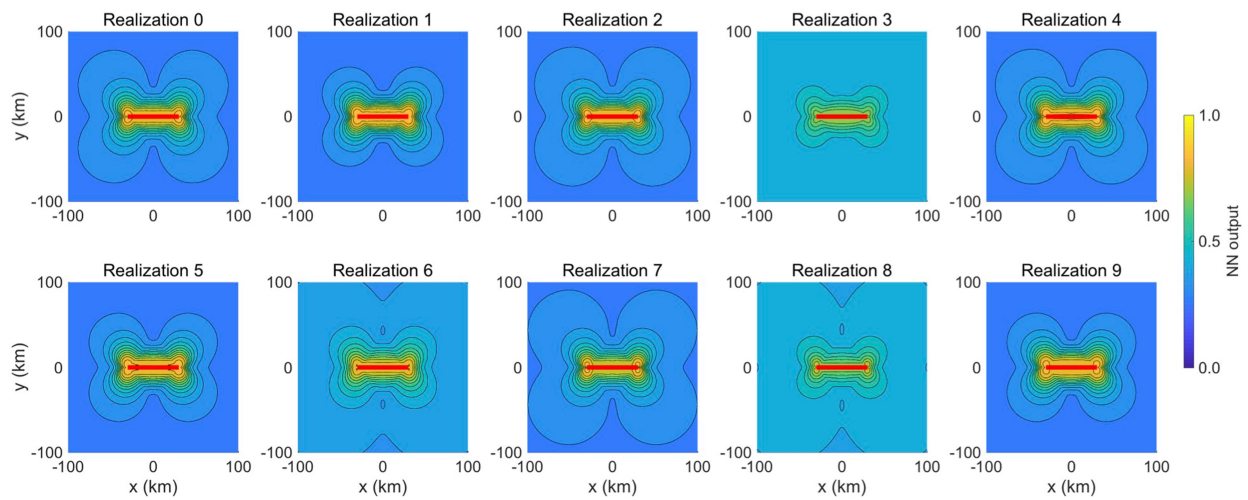
Extended Data Fig. 4 | ROC curves associated with realization 6 of the datasets. a–d, Curves incorporate grid cells down to a depth of 50 km. **e–h,** Curves including grid cells down to 5 km beyond the maximum depth of each slip distribution. Thus, the neural network in **d** is trained

and evaluated on a version of dataset realization 6 (Supplementary Table 2) that incorporates grid cells down to a depth of 50 km, whereas that in **h** is trained and evaluated on the same realizations of slip distributions, but incorporating only grid cells down to 5 km below each slip distribution.



Extended Data Fig. 5 | Forward predictions of the neural networks from each realization of the training dataset, incorporating all grid cells down to 50 km. Each panel is analogous to Fig. 2h, but uses one of

ten distinct neural networks trained on one of ten different realizations of the training dataset (Supplementary Table 2). See Methods for further discussion.



Extended Data Fig. 6 | Forward predictions of the neural networks from each realization of the training dataset, incorporating grid cells down to 5 km beyond the depth of each slip distribution. Each panel is

analogous to Fig. 2h, but uses one of ten distinct neural networks trained on one of ten different realizations of the training dataset (Supplementary Table 2). See Methods for further discussion.

Extended Data Table 1 | Comparison of physical metrics to the neural network for an idealized case

Quantity	Symbol	Evaluation	%VE
nearest distance	r	$r = \min(\sqrt{(x - x_f)^2 + (y - y_f)^2})$	46%
maximum shear	$\Delta\tau_{\max}(\chi)$	$\Delta\tau_{\max}(\chi) = \chi_1 - \chi_3 /2$	98%
1 st invariant	$\Delta I_1(\chi)$	$\Delta I_1(\chi) = \chi_1 + \chi_2 + \chi_3$	66%
2 nd invariant	$\Delta I_2(\chi)$	$\Delta I_2(\chi) = \chi_1\chi_2 + \chi_2\chi_3 + \chi_1\chi_3$	96%
von-Mises criteria	$\sqrt{3\Delta J_2}$	$\sqrt{3\Delta J_2} = \sqrt{\Delta I_1^2(\sigma) - 3\Delta I_2(\sigma)}$	98%
3 rd invariant	$\Delta I_3(\chi)$	$\Delta I_3(\chi) = \chi_1\chi_2\chi_3$	84%
Coulomb failure criteria $\mu = 0.0, 0.2, 0.4, 0.6, 0.8$	$\Delta CFS(\chi, \mu)$	$\Delta CFS(\chi, \mu) = (\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\parallel - \mu(\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\perp$	89%
Coulomb failure criteria normal only, $\mu = 0.4$	$\Delta CFS_n(\chi)$	$\Delta CFS_n(\chi) = -\mu(\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\perp$	59%
Coulomb failure criteria total shear	$\Delta CFS_\tau(\chi)$	$\Delta CFS_\tau(\chi) = (\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\parallel + (\mathbf{n}_\perp \cdot \chi) \cdot (\mathbf{n}_\parallel \times \mathbf{n}_\perp) $	95%
Coulomb failure criteria total, $\mu = 0.4$	$\Delta CFS_{\text{total}}(\chi, \mu)$	$\Delta CFS_{\text{total}}(\chi, \mu) = (\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\parallel + (\mathbf{n}_\perp \cdot \chi) \cdot (\mathbf{n}_\parallel \times \mathbf{n}_\perp) - \mu(\mathbf{n}_\perp \cdot \chi) \cdot \mathbf{n}_\perp$	93%
Sum of magnitudes of stress components	$m(\Delta\chi)$	$m(\Delta\chi) = \Delta\chi_{xx} + \Delta\chi_{yy} + \Delta\chi_{zz} + \Delta\chi_{xy} + \Delta\chi_{xz} + \Delta\chi_{yz} $	>99%

χ represents either the full (σ) or deviatoric (σ') stress-change tensor, χ_i are the corresponding eigenvalues, x_f and y_f are the x and y locations of the fault plane, respectively, and \mathbf{n}_\perp and \mathbf{n}_\parallel are the unit vectors perpendicular to the average orientation of the mainshock fault plane and parallel to the mean mainshock slip direction, respectively. %VE is the proportion of the variance in the strike-averaged neural-network forecast for the idealized strike-slip case (Fig. 2) that is explained by each strike-averaged physical metric. We include the largest %VE for each metric. For Coulomb failure stress change, the largest %VE corresponds to the magnitude of the Coulomb failure stress change associated with the full stress-change tensor $|\Delta CFS(\sigma, \mu = 0.0)|$. See Methods for details.

Extended Data Table 2 | Summary of results for ten realizations of the training and test datasets

Grid cells down to 50 km depth				
	$AUC_{\Delta CFS(\mu=0.4)}$	$AUC_{\Delta \tau_{max}}$	$AUC_{von-Mises}$	AUC_{NN}
Realization 0	0.5810	0.8147	0.8137	0.8185
Realization 1	0.6037	0.8104	0.8092	0.8127
Realization 2	0.6031	0.8067	0.8056	0.8123
Realization 3	0.6056	0.8174	0.8164	0.8224
Realization 4	0.5818	0.8183	0.8173	0.8200
Realization 5	0.5951	0.8083	0.8073	0.8128
Realization 6	0.6264	0.7985	0.7973	0.8002
Realization 7	0.5774	0.8202	0.8192	0.8242
Realization 8	0.6025	0.8214	0.8205	0.8258
Realization 9	0.5748	0.8161	0.8151	0.8186
Grid cells down to 5 km deeper than the depth of each slip distribution				
	$AUC_{\Delta CFS(\mu=0.4)}$	$AUC_{\Delta \tau_{max}}$	$AUC_{von-Mises}$	AUC_{NN}
Realization 0	0.5759	0.8016	0.8009	0.7982
Realization 1	0.5822	0.7937	0.7927	0.7946
Realization 2	0.5825	0.7946	0.7938	0.7964
Realization 3	0.6053	0.8073	0.8067	0.8095
Realization 4	0.5775	0.8054	0.8047	0.8066
Realization 5	0.5757	0.7937	0.7930	0.7941
Realization 6	0.5848	0.7817	0.7807	0.7828
Realization 7	0.5759	0.8087	0.8079	0.8096
Realization 8	0.6047	0.8110	0.8104	0.8133
Realization 9	0.5721	0.8018	0.8010	0.8000

The top half of the table displays results based on realizations that incorporate grid cells down to 50 km and the bottom half displays results based on realizations that incorporate grid cells down to 5 km below the maximum depth of each slip distribution (see Supplementary Table 2).

Fitness benefits and emergent division of labour at the onset of group living

Y. Ulrich^{1,2*}, J. Saragosti¹, C. K. Tokita³, C. E. Tarnita³ & D. J. C. Kronauer^{1*}

The initial fitness benefits of group living are considered to be the greatest hurdle to the evolution of sociality¹, and evolutionary theory predicts that these benefits need to arise at very small group sizes². Such benefits are thought to emerge partly from scaling effects that increase efficiency as group size increases^{3–5}. In social insects and other taxa, the benefits of group living have been proposed to stem from division of labour^{5–8}, which is characterized by between-individual variability and within-individual consistency (specialization) in task performance. However, at the onset of sociality groups were probably small and composed of similar individuals with potentially redundant—rather than complementary—function¹. Self-organization theory suggests that division of labour can emerge even in relatively small, simple groups^{9,10}. However, empirical data on the effects of group size on division of labour and on fitness remain equivocal⁶. Here we use long-term automated behavioural tracking in clonal ant colonies, combined with mathematical modelling, to show that increases in the size of social groups can generate division of labour among extremely similar workers, in groups as small as six individuals. These early effects on behaviour were associated with large increases in homeostasis—the maintenance of stable conditions in the colony¹¹—and per capita fitness. Our model suggests that increases in homeostasis are primarily driven by increases in group size itself, and to a smaller extent by a higher division of labour. Our results indicate that division of labour, increased homeostasis and higher fitness can emerge naturally in social groups that are small and homogeneous, and that scaling effects associated with increasing group size can thus promote social cohesion at the incipient stages of group living.

Quantifying the effects of group size on fitness and division of labour (DOL) requires the ability to precisely manipulate group size, monitor individual behaviour within groups and accurately measure fitness in controlled conditions. Crucially, group size must be controlled independently from factors such as colony genetic or age structure, which often co-vary with group size and which can affect fitness and DOL^{11,12}. To overcome these challenges, we use the clonal raider ant *Ooceraea biroi*, which combines the rich social biology of ants with unprecedented experimental amenability. This species displays an unusually simple social organization: colonies have no queens, and consist of genetically identical, monomorphic, totipotent workers that reproduce clonally and synchronously, and emerge in discrete age cohorts¹³. This provides maximal experimental control over group size and the genetic and demographic structure of colonies. Synchronized reproduction drives stereotypical colony cycles, in which colonies alternate between reproductive and brood-care phases, corresponding to the absence and presence of larvae. During the reproductive phase, all ants remain inside the nest and lay eggs. During the brood-care phase, the ants attend to the larvae inside the nest but also leave the nest—for example, to forage and to dispose of waste¹⁴.

We monitored the behaviour and fitness of colonies containing between 1 and 16 ants matched for genotype and age over at least one

colony cycle (Methods). Thus, within- and between-colony variation in genotype and age were minimal and could be ruled out as sources of variation in behaviour and fitness. Workload was standardized using a fixed initial larvae-to-workers ratio (1:1) across group sizes. The experiment was performed with 112 colonies from two clonal genotypes, A and B (previously labelled as MLL1 and MLL4¹⁵). Experiments started in the brood-care phase with workers and young larvae, and ended when all larvae in all colonies had either developed into adults or died. Colonies were kept in Petri dishes with no brood chamber (Fig. 1a); the ants freely chose a location to place their brood pile (henceforth, ‘the nest’). We analysed behaviour using custom automated image acquisition (7–9 frames per hour over 39–41 days) and analysis tools (Fig. 1a, Extended Data Fig. 1, Methods).

Because work in insect societies is spatially organized^{16,17} (for example, foraging and waste disposal occur away from the nest, whereas nursing occurs at the nest), individual behaviour can be described in terms of spatial location¹⁸. This is commonly done by assigning individuals to discrete behavioural groups on the basis of manually acquired spatial data¹³. Although the acquisition of spatial data has greatly improved with automated behavioural tracking^{18–20}, individuals are often still clustered into discrete behavioural groups¹⁸. However, in many systems—especially those without morphological castes—individual behaviour is continuously distributed^{21,22}. We therefore analysed behaviour non-parametrically from continuous spatial data, avoiding assumptions about the statistical distribution of individual behaviour. The spatial distribution of each ant was measured as the two-dimensional root-mean-square deviation (r.m.s.d.) of its *x* and *y* coordinates; that is, the spread of these coordinates around their centre of mass, throughout the brood-care phase (Fig. 1b, Extended Data Fig. 2a, Methods). The r.m.s.d. of an ant captures its tendency to explore the arena—that is, its lack of spatial fidelity (Fig. 1c)—and strongly correlates with its mean distance to the nest (Extended Data Fig. 3). The r.m.s.d. value is therefore a biologically meaningful metric that reflects the propensity to perform tasks away from the nest (for example, foraging) rather than at the nest (for example, nursing). In fact, the mean r.m.s.d. of a colony reflects its foraging activity: it increases when nutritional demand is elevated by increasing the larvae-to-workers ratio (Extended Data Fig. 4a, Supplementary Methods). As expected in this system, individuals varied in r.m.s.d., but did not cluster into discrete behavioural groups (Fig. 1b).

If DOL increases with group size, larger colonies are expected to show (1) higher behavioural variation between colony members and (2) higher individual behavioural consistency, or specialization, over time. Although not always independent from each other, these measures reflect distinct facets of DOL. Behavioural variation, computed as the standard deviation across r.m.s.d. values of ants from the same colony, increased with group size (Fig. 1d, Extended Data Fig. 5a), with small colonies (sizes 2–8) displaying less behavioural variation than larger colonies (sizes 12–16) (colony or group size refers to the number of individual workers and larvae added at the beginning of the experiment; for example, a colony of size 2 comprised 2 workers and 2 larvae). Short-term specialization was quantified as the r.m.s.d. rank

¹Laboratory of Social Evolution and Behavior, The Rockefeller University, New York, NY, USA. ²Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. ³Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. *e-mail: yuko.ulrich@gmail.com; dkronauer@rockefeller.edu

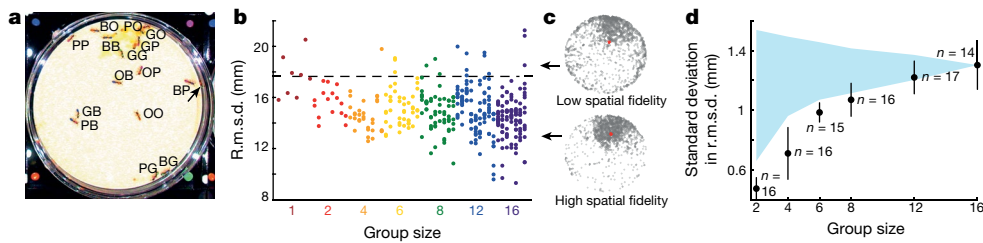


Fig. 1 | Behavioural variation as a function of group size. **a**, Example frame showing the result of automated ant detection: 15 correct colour-tag assignments (B, blue; G, green; O, orange; P, pink), 1 missed ant (arrow). **b**, Individual r.m.s.d. values for workers of genotype A. Ants from the same colony are vertically aligned. For the definition of r.m.s.d., see 'Behavioural data analysis' in Methods. Dashed line denotes the expected r.m.s.d. assuming a uniform distribution of an ant's positions. **c**, Spatial distribution of two ants from the same colony over the brood-care phase.

Red, centre of mass. Arrows point to the corresponding ants in **b**. Note that even workers with low spatial fidelity spend most of their time in the nest area. **d**, Behavioural variation increases with group size. Data for genotypes A and B are pooled. Black, standard deviation in r.m.s.d. per colony as a function of group size (mean \pm s.e.m.). Blue, 95% confidence intervals under the null hypothesis of no group-size effect on individual behaviour, generated by resampling individuals from colonies of size 16 (Extended Data Fig. 5a).

correlation between consecutive days, averaged over the first brood-care phase (Fig. 2a, Extended Data Fig. 5b–d); this captures the day-to-day behavioural consistency of group members relative to each other. Short-term specialization increased with group size and became significantly different from random at group size 6 (Fig. 2b). Long-term specialization, computed as the correlation between individual mean-r.m.s.d. ranks in the first and second brood-care phases (Extended Data Fig. 5e), was also found in colonies of six or more workers (Fig. 2c–f). Thus, DOL emerged at small group sizes, even in the absence of genetic and age variation, and increased as groups became larger.

We next used mathematical modelling to explore whether fixed response thresholds, the best theoretically studied self-organizing mechanism for DOL²³, could recapitulate these results. Individuals are assumed to respond to two task-related stimuli that reflect colony demand. A response occurs on the basis of innate, fixed

thresholds that determine the propensity of individual ants to perform a task given a certain stimulus level. For each task, individual thresholds are drawn from a normal distribution. The lower the stimulus intensity compared to an individual's threshold, the less likely it is to perform the task. The more sensitive this decision is to differences between stimulus intensity and threshold level, the more deterministic the threshold response (Methods). Individuals do not differ in their ability to perform tasks and there are no task-switching costs. When a task is performed, its stimulus level decreases; otherwise, it increases. Thus, across time, individuals divide their effort between tasks in various proportions, recapitulating a behavioural continuum that is similar to the experimental data. As long as there was some threshold variation among individuals, this simple model could robustly produce increased specialization with group size (measured as the slope between specialization values at group sizes 2 and 16) across a large parameter space,

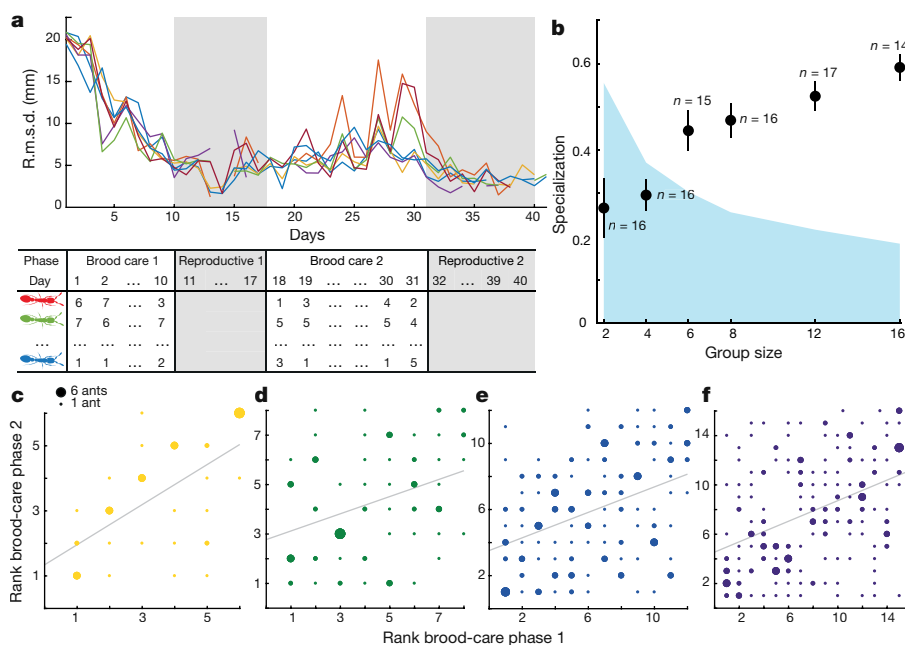


Fig. 2 | Specialization as a function of group size. **a**, Daily individual r.m.s.d. in one colony (size 8, genotype B). The matrix shows a subset of the corresponding daily r.m.s.d. ranks. **b**, Specialization (mean \pm s.e.m.) increases with group size. Black, mean day-to-day r.m.s.d.–rank correlation coefficients in the first brood-care phase as a function of group size. Positive values indicate a tendency for workers to maintain their behavioural rank across days, and 0 indicates that ranks are random. Blue, 95% confidence intervals generated by randomizing daily ranks (shuffling values along the columns of the matrix in **a**). **c**–**f**, Specialization persists across cycles in colonies of sizes 6–16. Grey lines, least-squares

fit. Spearman correlation between individual ranks over successive brood-care phases in colony size 6 (**c**) (r (degrees of freedom: 42) = 0.62, $P = 1.38 \times 10^{-5}$, 95% confidence interval (CI) -0.32 to 0.32), colony size 8 (**d**) (r (75) = 0.35, $P = 0.002$, 95% CI -0.23 to 0.24), colony size 12 (**e**) (r (160) = 0.40, $P = 1.31 \times 10^{-7}$, 95% CI -0.16 to 0.15), and colony size 16 (**f**) (r (209) = 0.44, $P = 3.42 \times 10^{-11}$, 95% CI -0.13 to 0.13). Circle diameter is proportional to the number of ants. Colonies of size 4 (r (32) = 0.08, $P = 0.68$, 95% CI: -0.42 to 0.42) are not shown. In **b**–**f**, data for genotypes A and B are pooled.

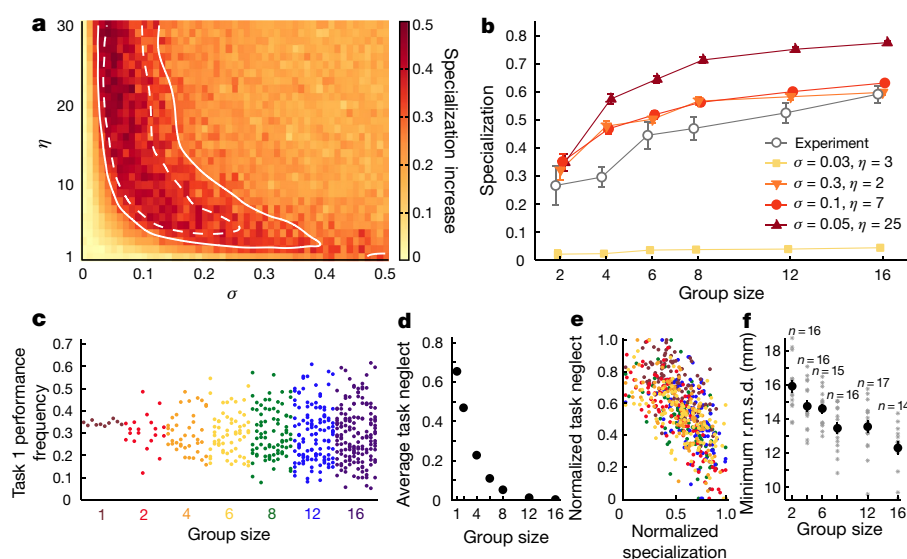


Fig. 3 | Results and predictions of the theoretical model. One hundred replicates were simulated per group size for each parameter combination. **a**, Increase in specialization between group size 2 and 16 (measured as the slope between these values) as a function of threshold stochasticity, η (higher η = a more deterministic threshold), and threshold variation, σ . The region between the solid (lower bound) and dashed (upper bound) white contours encompasses simulated slope values approximated to be within 10% of the experimental slope. **b**, Specialization (mean \pm s.e.m.) as a function of group size for different parameter combinations of the theoretical model (colour curves) and for experimental data (grey curve). **c–e**, One set of parameters (shown in Extended Data Fig. 6) corresponding to the filled circle symbol in **b**. **c**, Performance frequency of task 1 in

simulated colonies of various sizes (10 example replicates shown per group size). Each point represents an ant; ants from the same colony are vertically aligned. **d**, Average task neglect (that is, the proportion of time during a simulation run in which a task went unperformed) across tasks. Points represent the average value (mean \pm s.e.m.) across all simulated colonies. **e**, Relationship between specialization and task neglect when controlling for group size. Each point represents one simulated colony; colonies are coloured by group size as in **c**. **f**, Minimum r.m.s.d. (mean \pm s.e.m.), an empirical proxy for task neglect, decreases with group size (log-likelihood ratio test: $\chi^2 = 57.79$, $P = 2.92 \times 10^{-14}$). Asterisks represent colony data. Data for genotypes A and B are pooled.

including regions with a close quantitative match to our empirical observations (Fig. 3a, b, Supplementary Methods). Behavioural variation also increased with group size (Fig. 3c, Extended Data Fig. 6a, b). Beyond the group sizes used in our experiments, the model did not predict major further increases in DOL (Extended Data Fig. 6b, c). To explore how DOL might further increase with group size, other mechanisms—for example, direct social interactions²⁴ or spatial arrangement of tasks²⁵—would need to be considered.

The theoretically predicted increase in DOL with group size was robust across different specialization metrics^{10,26} (Extended Data Fig. 6d). Of these, the rank correlation metric (Fig. 2b) produced the highest values, which suggests it might be more sensitive to rudimentary types of specialization. However, the other metrics also reveal important insights: for example, task consistency—which quantifies how infrequently individuals switch between tasks (Supplementary Methods)—increased with group size even in the absence of task-switching costs, which suggests that reduced task switching could be an early emergent property of group living.

The theoretical analysis further revealed that increasing group size leads to increased homeostasis by (i) stabilizing stimuli intensities and task performance frequencies over time (Extended Data Fig. 7a, b), and (ii) decreasing task neglect—that is, instances in which tasks are not performed by any ant (Fig. 3d). These effects were obtained by increasing group size alone, even in the absence of DOL—that is, when there was no threshold variation—as long as some other source of stochasticity reduced the likelihood of behavioural synchronization among individuals (Extended Data Fig. 7a–c, Supplementary Notes). However, when present, DOL further increased homeostasis by enhancing some (Fig. 3e, Extended Data Fig. 7d)—although not all (Extended Data Fig. 7e)—of these effects. Larger colonies were thus more homeostatic than smaller colonies but, at a given size, more-specialized colonies had higher homeostasis. Subsequent analyses of the experimental data revealed dampened temporal fluctuations in colony-level behaviour (Extended Data Fig. 8a), increases in colony-level spatial stability

(Extended Data Fig. 8b) and decreases in minimum r.m.s.d. (a proxy for task neglect; Methods, Fig. 3f, Extended Data Fig. 9) with increasing group size, consistent with model predictions. These theoretical and empirical findings point to increases in colony homeostasis with group size via temporally more stable levels of work demand (for example, larval hunger) and more consistent work performance (for example, fewer instances of unattended brood). Because colony homeostasis is considered to be a key determinant of colony performance¹¹, the above results suggest that colony fitness should increase with group size.

Empirically, increases in group size were indeed associated with steep increases in fitness (Fig. 4a, Extended Data Fig. 2b). Colony growth rate was negative for the smallest colonies but rapidly increased with group size and plateaued at around 1 (indicating a doubling of size) in colonies of sizes 12 and 16 (Fig. 4b), similar to values reported in colonies that are orders-of-magnitude larger²⁷. Differences in colony growth were partially due to unexpected effects on brood development: the time to eclosion was 11 days (45%) longer in the smallest colonies compared to the largest colonies (Fig. 4c). This effect could not be recapitulated by varying larvae number alone (Extended Data Fig. 4b) and therefore probably arose from more efficient brood care in larger colonies. Because *O. biroi* colony cycles are controlled by the brood^{14,15}, the different times to eclosion suggest that small colonies had prolonged cycles. In fact, some large colonies had produced two cohorts of workers by the time small colonies had produced one (Extended Data Fig. 2b).

Control experiments and further analyses confirmed that none of our results were confounded by ant tagging, or by variation in ant density or morphology (Extended Data Fig. 10, Supplementary Methods).

In conclusion, we find that DOL, increased homeostasis and higher fitness can emerge as a function of group size at the incipient stages of group living. Notably, the rudimentary and flexible DOL demonstrated here does not rely on well-known mechanisms such as morphological caste specialization or age polyethism¹², but instead on plastic behavioural responses to the social environment^{28–30}. Although our theoretical model shows that specialization alone can increase group

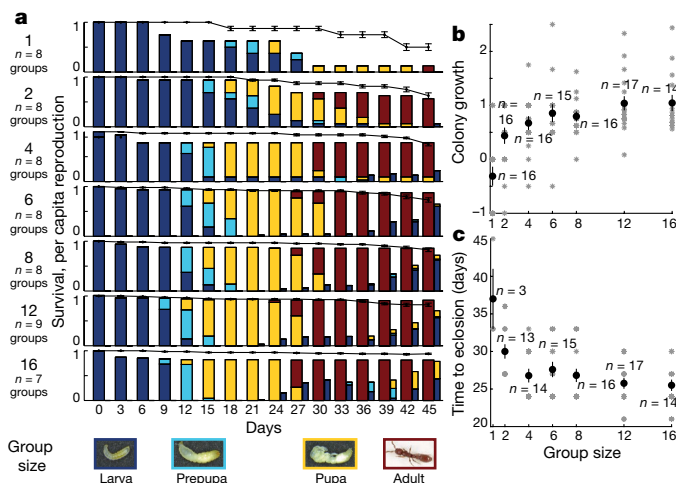


Fig. 4 | Fitness increases with group size. a, The dynamics of brood development as a function of group size in genotype A. Mean proportion of the brood in successive developmental stages (colours); the transition from larvae to pre-pupae marks the end of feeding and the switch from brood-care phase to reproductive phase. Wide and narrow bars indicate first and second brood generations, respectively. Black line, worker survival (mean ± s.e.m.). **b**, Colony growth (mean ± s.e.m.)—calculated as $(W_{end} - W_{start})/W_{start}$ in which W_{end} and W_{start} are the number of live workers at the end and start of the experiment, respectively—increases with group size (log-likelihood ratio test: $\chi^2 = 34.11$, $P = 5.22 \times 10^{-9}$). **c**, The time to eclosion (mean ± s.e.m.) decreases with group size (log-likelihood ratio test: $\chi^2 = 47.92$, $P = 4.44 \times 10^{-12}$). Sample sizes indicate the number of colonies in which at least one larva reached adulthood. In **b**, **c**, asterisks represent colony data, and data for genotypes A and B are pooled.

performance, future work is required to explore whether this direct link between DOL and fitness can be recapitulated experimentally. The scaling effects observed in our experiments and simulations provide a simple mechanism that could, along with other forces, promote social cohesion and provide an evolutionary stepping-stone towards more complex forms of social organization, such as those with morphologically differentiated queen and worker castes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0422-6>.

Received: 23 November 2017; Accepted: 27 June 2018;

Published online 22 August 2018.

- Queller, D. C. Cooperators since life began. *Q. Rev. Biol.* **72**, 184–188 (1997).
- Nowak, M. A., Tarnita, C. E. & Wilson, E. O. The evolution of eusociality. *Nature* **466**, 1057–1062 (2010).
- Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J. & Couzin, I. D. Emergent sensing of complex environments by mobile animal groups. *Science* **339**, 574–576 (2013).
- Morand-Ferron, J. & Quinn, J. L. Larger groups of passerines are more efficient problem solvers in the wild. *Proc. Natl Acad. Sci. USA* **108**, 15898–15903 (2011).
- Waters, J. S., Holbrook, C. T., Fewell, J. H. & Harrison, J. F. Allometric scaling of metabolism, growth, and activity in whole colonies of the seed-harvester ant *Pogonomyrmex californicus*. *Am. Nat.* **176**, 501–510 (2010).
- Dornhaus, A., Powell, S. & Bengtson, S. Group size and its effects on collective organization. *Annu. Rev. Entomol.* **57**, 123–141 (2012).
- Brahma, A., Mandal, S. & Gadagkar, R. Emergence of cooperation and division of labor in the primitively eusocial wasp *Ropalidia marginata*. *Proc. Natl Acad. Sci. USA* **115**, 756–761 (2018).
- Fewell, J. H. & Harrison, J. F. Scaling of work and energy use in social insect colonies. *Behav. Ecol. Sociobiol.* **70**, 1047–1061 (2016).
- Jeanson, R., Fewell, J. H., Gorelick, R. & Bertram, S. M. Emergence of increased division of labor as a function of group size. *Behav. Ecol. Sociobiol.* **62**, 289–298 (2007).
- Gautrais, J., Theraulaz, G., Deneubourg, J. L. & Anderson, C. Emergent polyethism as a consequence of increased colony size in insect societies. *J. Theor. Biol.* **215**, 363–373 (2002).

- Oldroyd, B. P. & Fewell, J. H. Genetic diversity promotes homeostasis in insect colonies. *Trends Ecol. Evol.* **22**, 408–413 (2007).
- Jeanson, R. & Weidenmüller, A. Interindividual variability in social insects - proximate causes and ultimate consequences. *Biol. Rev. Camb. Philos. Soc.* **89**, 671–687 (2014).
- Ravary, F. & Jaisson, P. Absence of individual sterility in thelytokous colonies of the ant *Cerapachys biroi* Forel (Formicidae, Cerapachyinae). *Insectes Soc.* **51**, 67–73 (2004).
- Ravary, F., Jahyny, B. & Jaisson, P. Brood stimulation controls the phasic reproductive cycle of the parthenogenetic ant *Cerapachys biroi*. *Insectes Soc.* **53**, 20–26 (2006).
- Oxley, P. R. et al. The genome of the clonal raider ant *Cerapachys biroi*. *Curr. Biol.* **24**, 451–458 (2014).
- Sendova-Franks, A. B. & Franks, N. R. Spatial relationships within nests of the ant *Leptothorax unifasciatus* (Latr) and their implications for the division of labor. *Anim. Behav.* **50**, 121–136 (1995).
- Gordon, D. M. Dynamics of task switching in harvester ants. *Anim. Behav.* **38**, 194–204 (1989).
- Mersch, D. P., Crespi, A. & Keller, L. Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science* **340**, 1090–1093 (2013).
- Heyman, Y., Shental, N., Brandis, A., Hefetz, A. & Feinerman, O. Ants regulate colony spatial organization using multiple chemical road-signs. *Nat. Commun.* **8**, 15414 (2017).
- Crall, J. D. et al. Spatial fidelity of workers predicts collective response to disturbance in a social insect. *Nat. Commun.* **9**, 1201 (2018).
- Weidenmüller, A. The control of nest climate in bumblebee (*Bombus terrestris*) colonies: interindividual variability and self reinforcement in fanning response. *Behav. Ecol.* **15**, 120–128 (2004).
- Campos, D., Bartumeus, F., Méndez, V., Andrade, J. S. Jr & Espadaler, X. Variability in individual activity bursts improves ant foraging success. *J. R. Soc. Interface* **13**, 20160856 (2016).
- Bonabeau, E., Theraulaz, G. & Deneubourg, J.-L. Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies. *Proc. R. Soc. Lond. B* **263**, 1565–1569 (1996).
- Pacala, S. W., Gordon, D. M. & Godfray, H. C. J. Effects of social group size on information transfer and task allocation. *Evol. Ecol.* **10**, 127–165 (1996).
- Franks, N. R. & Tofts, C. Foraging for work: how tasks allocate workers. *Anim. Behav.* **48**, 470–472 (1994).
- Gorelick, R., Bertram, S. M., Killeen, P. R. & Fewell, J. H. Normalized mutual entropy in biology: quantifying division of labor. *Am. Nat.* **164**, 677–682 (2004).
- Teseo, S., Châline, N., Jaisson, P. & Kronauer, D. J. C. Epistasis between adults and larvae underlies caste fate and fitness in a clonal ant. *Nat. Commun.* **5**, 3363 (2014).
- Crall, J. D. et al. Social context modulates idiosyncrasy of behaviour in the gregarious cockroach *Blaberus discoidalis*. *Anim. Behav.* **111**, 297–305 (2016).
- Freund, J. et al. Emergence of individuality in genetically identical mice. *Science* **340**, 756–759 (2013).
- Holbrook, C. T., Kukuk, P. F. & Fewell, J. H. Increased group size promotes task specialization in a normally solitary halictine bee. *Behaviour* **150**, 1449–1466 (2013).

Acknowledgements We thank A. Gal for advice on data analysis, O. Feinerman and M. Liu for contributions to the tracking algorithms, S. Leibler, Z. Frentz, and D. Jordan for helpful discussions. This work was supported by grant 1DP2GM105454-01 from the NIH, a Searle Scholar Award, a Klingenstein-Simons Fellowship Award in the Neurosciences, and a Pew Biomedical Scholar Award to D.J.C.K.; Swiss National Science Foundation Early Postdoc. Mobility (PBEZP3-140156) and Advanced Postdoc. Mobility (P300P3-147900) fellowships, and a Rockefeller University Women & Science fellowship to Y.U.; a Kravis Fellowship to J.S.; the National Science Foundation Graduate Research Fellowship under Grant No. DGE1656466 to C.K.T. This is Clonal Raider Ant Project paper number 8.

Reviewer information Nature thanks J. O'Dwyer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions This study was conceived by Y.U. and D.J.C.K. Experiments were designed by Y.U. and D.J.C.K. Tracking hardware and software were developed by J.S. and Y.U. Empirical data were analysed by Y.U. Theoretical modelling was performed by C.K.T. and C.E.T. Computational modelling was performed by C.K.T. Y.U. and D.J.C.K. drafted the manuscript. D.J.C.K. supervised the project. All authors revised the manuscript and approved the final version for publication.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0422-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0422-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.U. or D.J.C.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. Individual ants were assigned to experimental treatments (colonies of different size) randomly. The investigators were not blinded to allocation during experiments and outcome assessment.

Experimental design. Experimental colonies were composed of age-matched, one-cycle-old workers (44 and 34 days old for genotypes A and B, respectively; A colonies have slower cycles than B colonies on average) and 4-day-old larvae in airtight Petri dishes (5 cm in diameter, corresponding to about 25 ant body-lengths) with a plaster of Paris floor. All workers and larvae within an experiment—including replicate colonies of all group sizes—were clonally related and sourced from the same stock colony. All workers within an experiment were also collected from the same cohort and had eclosed within a day of each other (owing to the synchronized reproduction of *O. biroi*). From the time they were collected (1–3 days after eclosion) until the start of the experiment, workers were kept together in a box and allowed to go through a full colony cycle. Thus, all workers within an experiment experienced the same environment as larvae and as adults. However, we cannot exclude the possibility that small differences in individual experience occurred even in this common environment before the start of the experiment. All workers were tagged with colour marks on the thorax and gaster using oil-paint markers (uni Paint Markers PX-20 and PX-21). Experimental colonies contained 1, 2, 4, 6, 8, 12 or 16 workers and a matching number of larvae. This 1:1 larvae-to-workers ratio corresponds to the estimated ratio found in a typical (that is, large and healthy) laboratory stock colony in the brood-care phase. The experiment was conducted using two distinct genotypes, A and B¹⁵. Between 7 and 9 replicate colonies were used for each group size and genotype, giving a total of 112 colonies. *O. biroi* is myrmecophilous and colonies were fed live pupae of fire ant (*Solenopsis invicta*) minor workers. These prey items are small enough to be transported by a single *O. biroi* worker, so small colonies were not disproportionately penalized by the feeding regime.

The experiments took place in a climate room at 25 °C and 75% relative humidity under constant light (*O. biroi* is blind and its behaviour is not affected by light). Every three days, we cleaned and watered the plaster, added one prey item per live larva at a random location within the Petri dish, and recorded adult survival as well as brood survival and development under a stereomicroscope in all colonies (except for eggs, which cannot be counted without substantially disturbing the colony). The experiments ended when all larvae within an experiment had either developed into adult workers or died. Two colonies (of sizes 6 and 16, genotype B) were excluded from all analyses owing to setup errors (incorrect number of workers or larvae at the beginning of the experiment). Note that although we controlled the number of workers and larvae at the beginning of the experiment, these numbers then changed throughout the experiment as workers died and reproduced, and as the brood died or developed into adults.

Image acquisition and ant detection. Behavioural data were acquired using an automated scan-sampling approach, in which a picture of each colony was taken at regular intervals throughout the experiment. For this purpose, we designed and built a setup comprising 28 webcams (Logitech B910 or C910) and controlled LED lighting. Each webcam acquired images of four colonies, and the position of colonies within the setup was randomized. This resulted in 7,976 and 6,429 frames per colony over 39 and 41 days for genotypes A and B, respectively. The difference in overall frame rate between the two experiments stems in part from the variability in image acquisition speed of the computers used to control the webcams (median interval between frames: 420 s for genotype A, 525 s for genotype B), and in part from an approximately 35-h-long scanning interruption in the genotype B experiment. This interruption occurred in the reproductive phase of most colonies (51 out of 56). Because behavioural analyses were conducted on data collected in the brood-care phase only, this interruption did not affect our results and conclusions. Fitness monitoring outlasted behavioural data acquisition by 6 days in genotype A to allow the last callow workers to eclose.

Within-image variation in lighting and hue was corrected by dividing each frame's RGB values by those of an image of a uniformly grey surface taken with the same camera immediately before the start of the experiment (Extended Data Fig. 1a). After manual selection of the image region corresponding to the plaster arena, a Bayesian classifier was used to assign to each pixel a probability of belonging to each of the following eight colour categories: plaster, ant cuticle, shadow, food and colour tags (pink, orange, blue and green) (Extended Data Fig. 1b). Size and colour-probability thresholds were used to detect candidate regions corresponding to ants carrying colour tags (Extended Data Fig. 1c). Candidate ants were oriented on the basis of the relative position of cuticle and tag colour probability maxima along the main axis of each region (for example, given a candidate ant carrying a blue and a green tag, the blue tag can be assigned to the thorax and the green tag to the gaster if pixels with high cuticle colour probability, corresponding to the ant's head, can be found next to the blue but not the green tag) (Extended Data Fig. 1d). Candidate ants were assigned a final ID using Munkres' variant of the Hungarian

assignment algorithm (Extended Data Fig. 1e). Performance of the automated assignments was assessed by comparison with manual assignments for 280 frames selected randomly throughout the first brood-care phase and across colonies in the genotype B experiment. On average, the ant identification algorithms correctly identified 77.1% of the ants that could be manually identified (that is, 22.9% of ants were missed). Of all the automated assignments, 94.4% were correct (that is, 5.6% assigned the wrong ID).

Additionally, we performed manual assignments at a higher frequency (every 10 frames) for one 16-worker colony and verified that individual behavioural traits computed from automated assignments correlated with the same traits computed from manual assignments. Individual behavioural r.m.s.d. values calculated from automated tracking data strongly correlated with the same values calculated from manual tracking (Extended Data Fig. 1f). Software for automated image acquisition and analysis was developed in MATLAB.

Behavioural data analysis. We restricted our behavioural analyses to the brood-care phase because worker locomotion, and thus our ability to detect inter-individual behavioural differences, is markedly reduced during the reproductive phase. For each colony, the brood-care phase started at the beginning of the experiment and ended when all larvae had either reached the non-feeding pre-pupal stage (that is, ejected their meconium) or died. The end of the brood-care phase was scored by visual inspection of images. Note that this definition of the brood-care phase—based solely on the brood developmental stage—is discrete and differs from that of a previous study³¹, in which the brood-care phase was characterized using both the development of larvae and the foraging activity of workers.

The spatial distribution of each ant throughout the brood-care phase was quantified as the two-dimensional r.m.s.d.

$$\text{r.m.s.d.} = \sqrt{\frac{\sum_i ((x_i - \bar{x})^2 + (y_i - \bar{y})^2)}{n}}$$

in which x_i and y_i are the coordinates of the focal ant in frame i , \bar{x} and \bar{y} are the coordinates of the centre of mass of the focal ant's overall spatial distribution in the considered time frame, and n is the number of frames in which the focal ant was detected. The r.m.s.d. is bounded between 0 and r , the radius of the Petri dish. Workers that spend a lot of time in the nest with the brood (for example, nursing the larvae) and little time performing outside tasks (foraging or waste disposal) have low r.m.s.d. values, whereas workers that spend comparatively more time away from the brood have higher r.m.s.d. values. In previous studies, workers displaying behaviour corresponding to low or high r.m.s.d. have been labelled 'nurses' and 'foragers', respectively. However, given the apparent continuous distribution of r.m.s.d. values across individuals in this study (Fig. 1b), we chose not to cluster individuals into discrete behavioural 'castes'.

For each colony, mean behaviour was computed as the average of individual r.m.s.d. values, and behavioural variability was computed as the standard deviation of individual r.m.s.d. values. Both metrics were then averaged across replicate colonies for each group size. Artefacts due to sampling effects are of particular concern for any experiment in which variation in group size is an experimental treatment. For this reason, whenever possible we compared group sizes using resampling and randomization approaches in addition to standard statistical tests (Supplementary Notes). To assess the significance of any effect of group size on behaviour while ruling out sampling effects, we simulated colonies of sizes 1 to 12 by randomly sampling 1 to 12 individuals (without replacement) from each colony of size 16 (Extended Data Fig. 5a). Mean behaviour and behavioural variability were calculated for each simulated colony and averaged across replicate colonies of a given size, as described above. This resampling procedure was repeated 1,000 times. Ninety-five per cent confidence intervals were generated for mean behaviour and behavioural variation for each group size separately using the same resampled data, to test which colony size had an observed behaviour that significantly differed from that of colonies of size 16.

To quantify specialization, we introduce a metric appropriate for use in very small colonies and on continuous behavioural data. Existing measures of specialization usually require discrete tasks to be defined and scored, which is generally done by a human observer. Task definition is subjective with respect to the nature and number of defined tasks, and task scoring is susceptible to inter-observer variation. Thus, instead of using existing task-based measures of specialization, our metric is based on continuous behavioural data (r.m.s.d.): colony-level behavioural consistency, or specialization, was defined for each colony as the correlation coefficient between individual r.m.s.d. ranks on consecutive days, averaged over the first brood-care phase (Extended Data Fig. 5b–d). Spearman rank correlations, rather than parametric correlations (for example, Pearson), were used because there was more variation in r.m.s.d. over time than across individuals (Fig. 2a). The timeframe of days was chosen to ensure that individual r.m.s.d. values were calculated on a sufficient number of detections (150–200) for each time interval. These mean rank-correlation coefficients were then compared across colonies of different sizes. This measure can

be used to describe specialization in very small colonies (starting at size 2). To assess significance for each group size, 95% confidence intervals for rank-correlation coefficients were generated by randomizing ranks on each day in each colony 1,000 times, based on the null hypothesis that worker behaviour is uncorrelated across successive days. We then tested whether individual behaviour was consistent over successive brood-care phases. To do this, we selected colonies that had a second brood-care phase (defined by the presence of a new cohort of larvae hatched from eggs laid by the workers during the first reproductive phase) for which at least four days of behavioural data were available. In these colonies, for each brood-care phase, workers were assigned a within-colony rank on the basis of their mean r.m.s.d. across days (Extended Data Fig. 5e). Long-term behavioural consistency was defined as a significant positive correlation between the individual ranks in each brood-care phase. To assess significance, 95% confidence intervals for rank-correlation coefficients were generated by randomizing ranks in each brood-care phase for each colony 1,000 times, based on the null hypothesis that worker behaviour is uncorrelated across successive brood-care phases. Correlations were computed for all workers within colonies of the same size. The analysis could not be performed for colonies of size 2 because too few colonies of this size had a second brood-care phase.

Finally, we investigated whether behavioural fluctuations and task neglect decreased with group size. To compute behavioural fluctuations, we calculated colony mean r.m.s.d. by averaging the daily individual r.m.s.d. values of all colony members, computed the fluctuations (that is, absolute differences) of this colony mean r.m.s.d. between successive days, averaged these differences across the first brood-care phase and compared these mean fluctuations across colonies of different sizes.

Task neglect was computed using a r.m.s.d.-based proxy indicative of the consistent performance of tasks taking place in the nest. This metric, the minimum r.m.s.d., is the r.m.s.d. value of the ant with the highest spatial fidelity to the nest in each colony. The lower the minimum r.m.s.d., the more likely it is that at least one ant is at the nest—that is, that the brood is not left unattended. Task neglect was also quantified as the proportion of times in which the brood was unattended (that is, when no worker was found in the nest), using a combination of manual and automated tracking. To this aim, the position of the brood pile was annotated manually (if it could be determined by eye from images) every three days for each colony (Extended Data Fig. 9a). The distance between individual ant positions (obtained from automated tracking) and the brood pile was calculated for each frame in the previous three days (that is, distances were calculated between ant positions on days 1–3 and the brood-pile position on day 3). An ant was considered to be in the nest if it was within 5 mm of the brood-pile contour. For each colony, ‘observed task neglect’ was defined as the proportion of frames of the brood-care phase in which no ant was found in the nest (that is, task neglect implicitly refers to nursing here) (Extended Data Fig. 9b). Because larger colonies have higher ant density, the probability that at least one ant is found in the nest (as in any other area of the Petri dish) could increase with group size in a trivial way, without any associated change in individual behaviour. To control for this, each manually annotated brood area was also rotated by 180° around the centre of the Petri dish to produce a control area in the box of the same shape and area as the brood area (Extended Data Fig. 9a), and the number of ants in that random area was counted as above to produce a measure of ‘expected task neglect’ under the null model of no behavioural change (Extended Data Fig. 9b). If task neglect decreases with group size, we expected the difference between observed and expected task neglect (or ‘effective task neglect’) to decrease with group size (Extended Data Fig. 9c).

For all behavioural analyses, ants were excluded from the dataset if they were detected in less than 30% of the frames acquired within the considered time frame (brood-care phase or day; for ants that died during the brood-care phase, the considered time frame was the portion of the brood-care phase preceding death). This is unlikely to have introduced a bias because low-r.m.s.d. and high-r.m.s.d. workers have similar detection probabilities (see sample sizes in Extended Data Fig. 1f).

Statistical analyses. The effects of group size (1, 2, 4, 6, 12 or 16), genotype (A or B) and their interaction on behaviour (mean r.m.s.d., standard deviation of r.m.s.d., behavioural consistency, behavioural fluctuations, minimum r.m.s.d. and task neglect) and fitness (colony growth and time to eclosion) were investigated using generalized linear models. Colonies of size 1 were excluded from the models of standard deviation of r.m.s.d., behavioural consistency and minimum r.m.s.d. because the corresponding values were constant at 0, undetermined and uninformative, respectively. When needed, response variables were transformed to satisfy model assumptions of normally distributed residuals (tested with a Wilk–Shapiro test) and homoscedasticity (tested with Levene’s test). We evaluated the significance of effects and their interaction by comparing pairs of nested models using χ^2 log-likelihood ratio tests following deletion of terms (starting with the interaction). Data from genotypes A and B were pooled whenever justified by the absence of a significant interaction term between the effect of genotype and the effect of group size. Statistical analyses were performed in R³². Full statistical results are presented as Supplementary Notes.

Theoretical model. First introduced to the social insect literature by Bonabeau et al.²³, fixed response thresholds have been a widely used approach to study the emergence of DOL in self-organized social systems. The model considers n individuals and assumes that there are two possible states for any given individual—active and inactive. Active individuals perform exactly one of m tasks at any moment in time (for simplicity, we assume that there are only two tasks, that is $m = 2$). Inactive individuals do not perform any task; they are considered to be in a rest state. An n by m binary matrix, $X_t = [x_{ij,t}]$, describes the activity and task state of each individual at a given time step t : if individual i is inactive, then all $x_{ij,t} = 0$ because an inactive individual performs no task; if individual i is active, then exactly one $x_{ij,t} = 1$ while all others are 0.

The model assumes that each task j has an associated stimulus, $s_{j,t}$, which signals the group-level demand for that task at time t . The change in the stimulus over discrete time can be modelled according to a previously published study³² as:

$$s_{j,t+1} = s_{j,t} + \delta_j - \alpha \frac{\sum_{i=1}^n x_{ij,t}}{n}$$

in which δ_j is the constant, task-specific stimulus increase rate per time step (that is, task demand rate), α is a scalar measuring task performance efficiency (assumed, for simplicity, to be the same for all individuals across all tasks), $\sum_{i=1}^n x_{ij,t}$ is the number of individuals performing task j at time t and n is the total number of individuals in the colony. The higher the α , the better the individuals are at performing the tasks. The higher the δ_j , the more demanding that task; for simplicity, all tasks are assumed to have the same demand rate, δ .

The model has four sources of stochasticity. First, each individual i has an internal threshold, θ_{ij} , for each task j . This is randomly drawn from a normal distribution with mean μ_j and normalized standard deviation σ_j , which is given as a proportion of μ_j (for example, $\sigma = 0.3$ indicates a standard deviation that is 30% of the mean). For simplicity, we assume that μ and σ are the same for all tasks. Second, inactive individuals are exposed to task stimuli randomly in a given time step, until they either commit to performing a given task and thus become active, or cycle through all stimuli without becoming active and thus remain inactive. Third, for each encountered stimulus, individuals determine whether to perform that task by evaluating the stimulus level relative to their corresponding internal threshold. The threshold response function that gives the probability $P_{ij,t}$ that individual i performs task j at time t is a sigmoid for which the steepness is determined by a parameter η to range from more deterministic to more stochastic³³:

$$P_{ij,t} = \frac{s_{j,t}^{\eta}}{s_{j,t}^{\eta} + \theta_{ij}^{\eta}}$$

For large values of η , we recover a deterministic behaviour, such that a task is only performed if the stimulus exceeds the threshold, and in that case it is always performed. In our simulations, we vary $1 \leq \eta \leq 30$ to capture this range of behaviours. Fourth, upon starting a task, an individual will continue performing that task until it spontaneously quits, with a constant quit probability, $\tau^{9,10,23,34}$. Active individuals do not evaluate task stimuli and do not switch between tasks; only inactive individuals evaluate stimuli and determine which task (if any) to start performing.

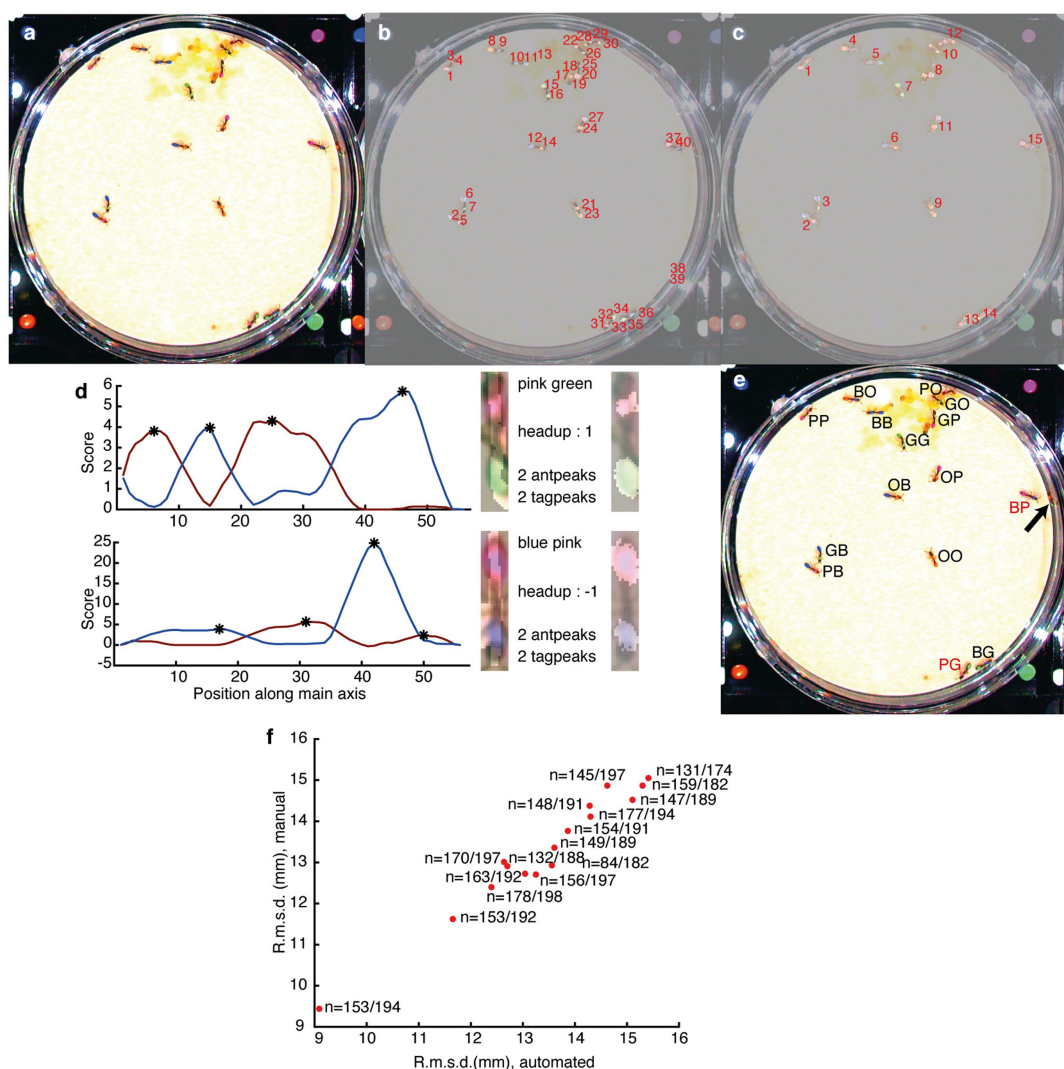
To analyse this model and—specifically—how each of the four sources of stochasticity affects the outcome, we started from a fully deterministic version and built in each one of the different sources of stochasticity independently (Extended Data Fig. 7, Supplementary Methods). All agent-based simulations and subsequent data analyses were conducted in R³².

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All behavioural tracking code is available at <https://doi.org/10.5281/zenodo.1211644>. All code for model simulations is available at <https://doi.org/10.5281/zenodo.1211231>.

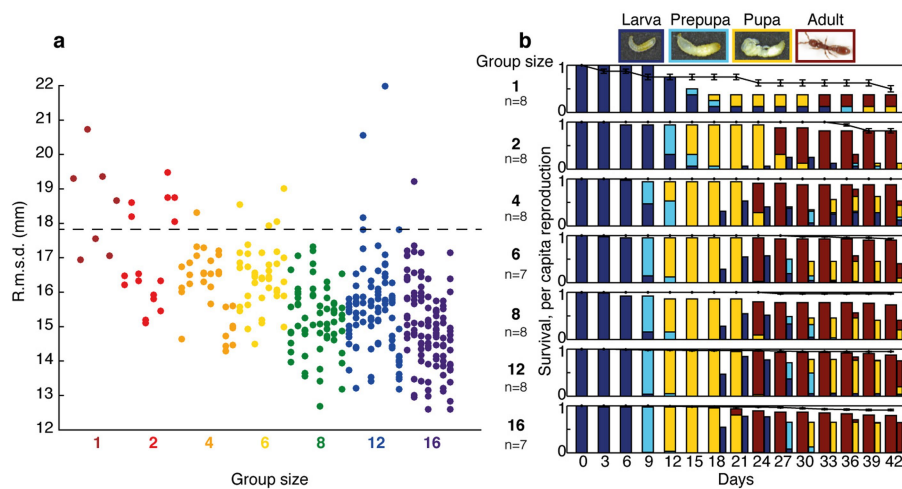
Data availability. All behavioural tracking data (x , y positions of individual ants in different frames) as well as colony summary statistics (behaviour and fitness) are available at <https://doi.org/10.5281/zenodo.1237867>. Any other data that support the findings of this study, such as processed data files used for statistical analyses, are available from the corresponding authors upon reasonable request.

1. Ravary, F. & Jaisson, P. The reproductive cycle of thelytokous colonies of *Cerapachys biroi* Forel (Formicidae, Cerapachyinae). *Insectes Soc.* **49**, 114–119 (2002).
2. R Core Team. *R: A Language and Environment for Statistical Computing* <http://www.R-project.org/> (R Foundation for Statistical Computing, Vienna, 2008).
3. Dodds, P. S. & Watts, D. J. Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.* **92**, 218701 (2004).
4. Bonabeau, E., Theraulaz, G. & Deneubourg, J.-L. Fixed response thresholds and the regulation of division of labor in insect societies. *Bull. Math. Biol.* **60**, 753–807 (1998).



Extended Data Fig. 1 | Ant detection algorithm. **a**, Example cropped frame showing one 16-worker colony after image correction. **b**, Colour-tag detection. The highlighted numbered zones are image regions containing pixels that were assigned a high probability for tag colours (green, blue, orange or pink) by a Bayesian classifier. **c**, Candidate ant detection. The highlighted numbered zones correspond to contiguous regions containing colour tags and pixels that were assigned a high probability for ant colour (that is, cuticle) by the classifier. **d**, Candidate ant orientation. Candidate ants are aligned using the segment connecting the two colour tags, and oriented (head down versus head up) on the basis of the relative position of cuticle- and tag-probability maxima (black stars on brown and blue

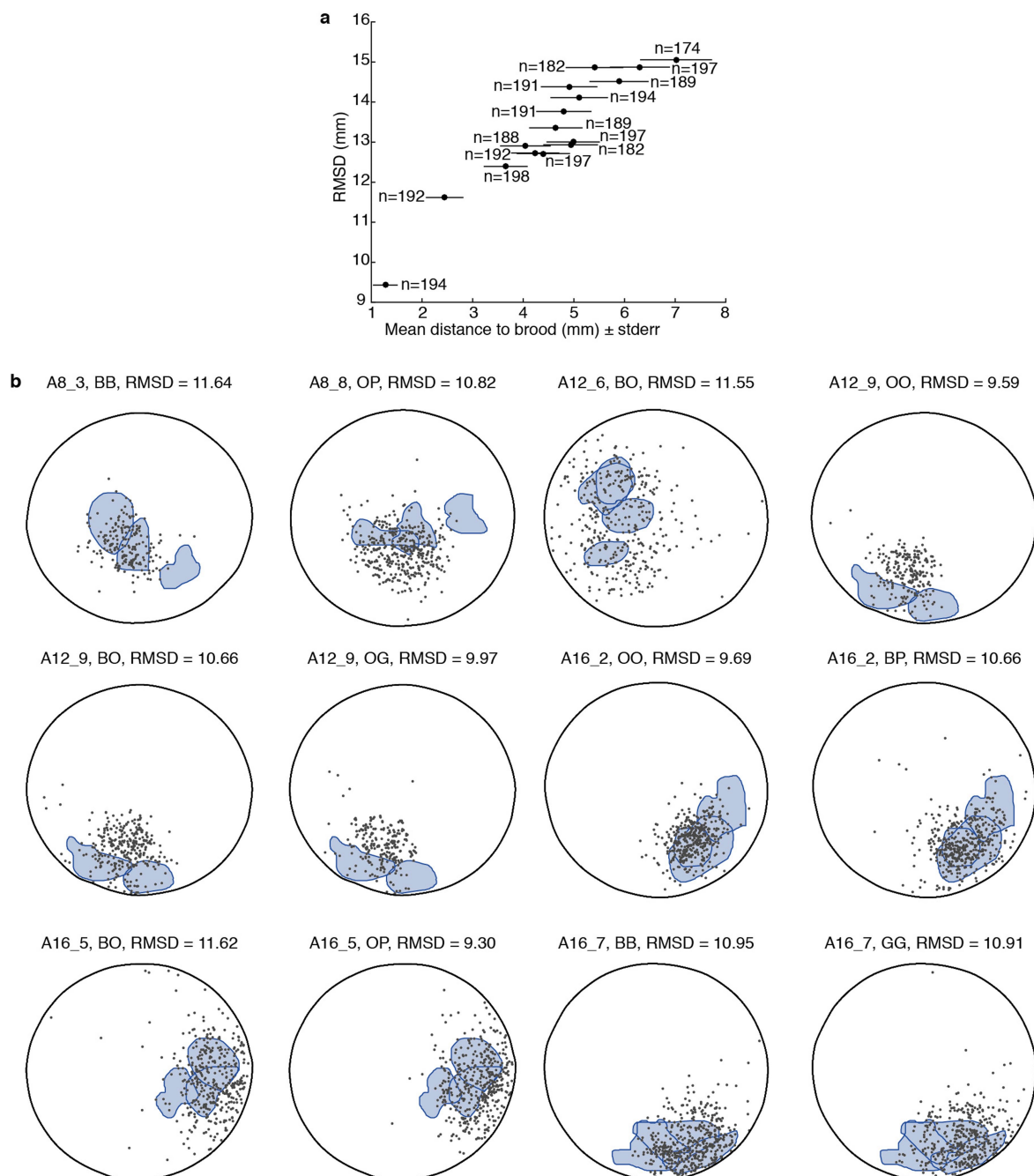
lines, respectively) along the main axis of each region. **e**, Final IDs after using Munkres' variant of the Hungarian assignment algorithm. Labels indicate colour IDs (thorax–abdomen; G, green; B, blue; O, orange; P, pink). Ants shown as examples in **d** are labelled in red. All assignments shown are correct, but one ant is missed (arrow). This panel is identical to Fig. 1a. **f**, Correlation between r.m.s.d. calculated from automated versus manual assignments for one 16-worker colony. The r.m.s.d. was computed from a subset of frames in the brood-care phase. $n = (\text{number of automated detections})/(\text{number of manual detections})$. Pearson's $r = 0.95$, $P < 0.001$.



Extended Data Fig. 2 | The r.m.s.d. and fitness in genotype B.

a, Individual r.m.s.d. values for all workers of genotype B. Ants from the same colony are vertically aligned. The dashed line represents the expected r.m.s.d. assuming a uniform distribution of an ant's positions. **b**, The dynamics of brood development as a function of group size in genotype B.

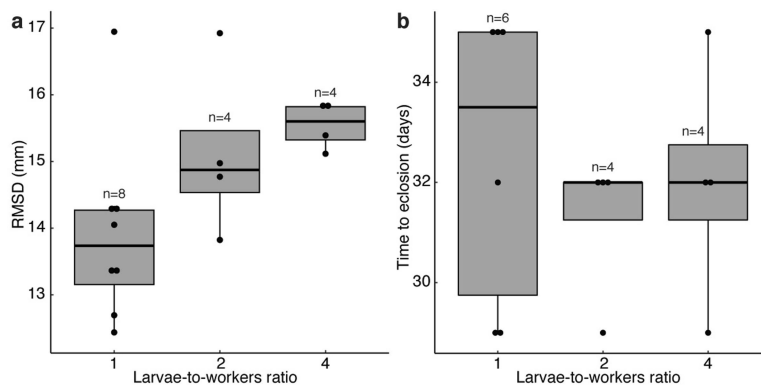
The proportion of the brood in successive developmental stages (colours) in colonies of sizes 1–16 is shown. Wide and narrow bars indicate first and second brood generations, respectively. Black line, worker survival (mean \pm s.e.m.).



Extended Data Fig. 3 | The r.m.s.d. and spatial fidelity to the nest.

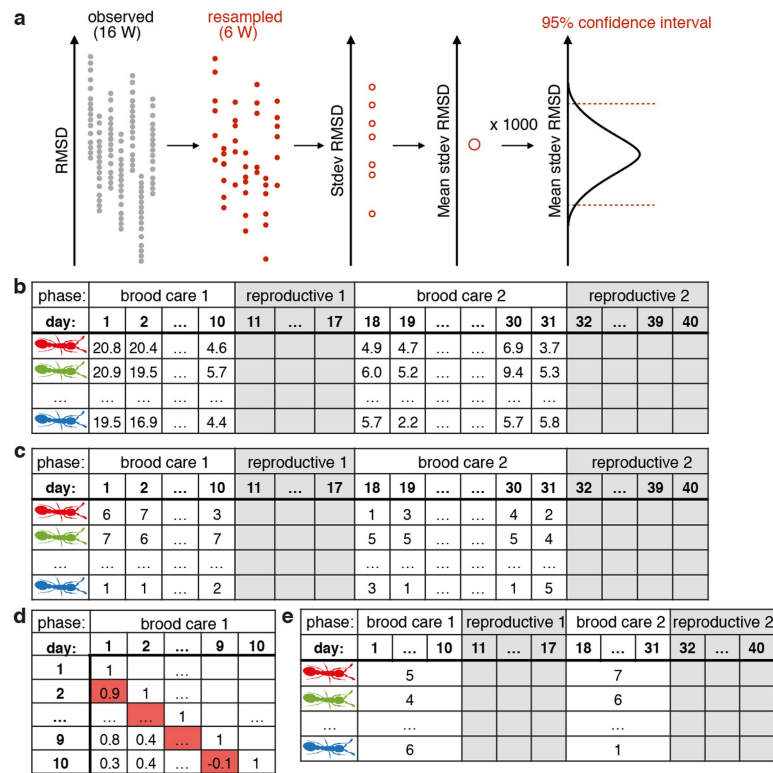
a, Correlation between individual r.m.s.d. and individual distance to the brood (mean \pm s.e.m.) over the first brood-care phase in one colony of 16 workers (Spearman's $r = 0.93$, $P = 0$; $n = 16$ ants). Behavioural traits are based on 209 manually tracked frames. Sample sizes (n) indicate the number of frames in which each ant was manually identified. Manual tracking was used here because the automated tracking algorithm does not allow us to locate the brood. **b**, Individuals with low r.m.s.d. (r.m.s.d. < 12 in Fig. 1b) have high spatial fidelity to the nest area. Each circle represents

the spatial distribution of an ant (grey dots) with respect to the brood pile (shaded blue areas) in the brood-care phase. Panel titles indicate colony identity (for example, A8_3 is the third replicate colony of genotype A and size 8), ant identity (for example, BO for blue–orange) and individual r.m.s.d. In each colony, the brood pile was manually annotated every three days (that is, if the brood-care phase lasted nine days, three brood piles zones were manually annotated; zones could overlap or not, depending on how much the brood pile moved).



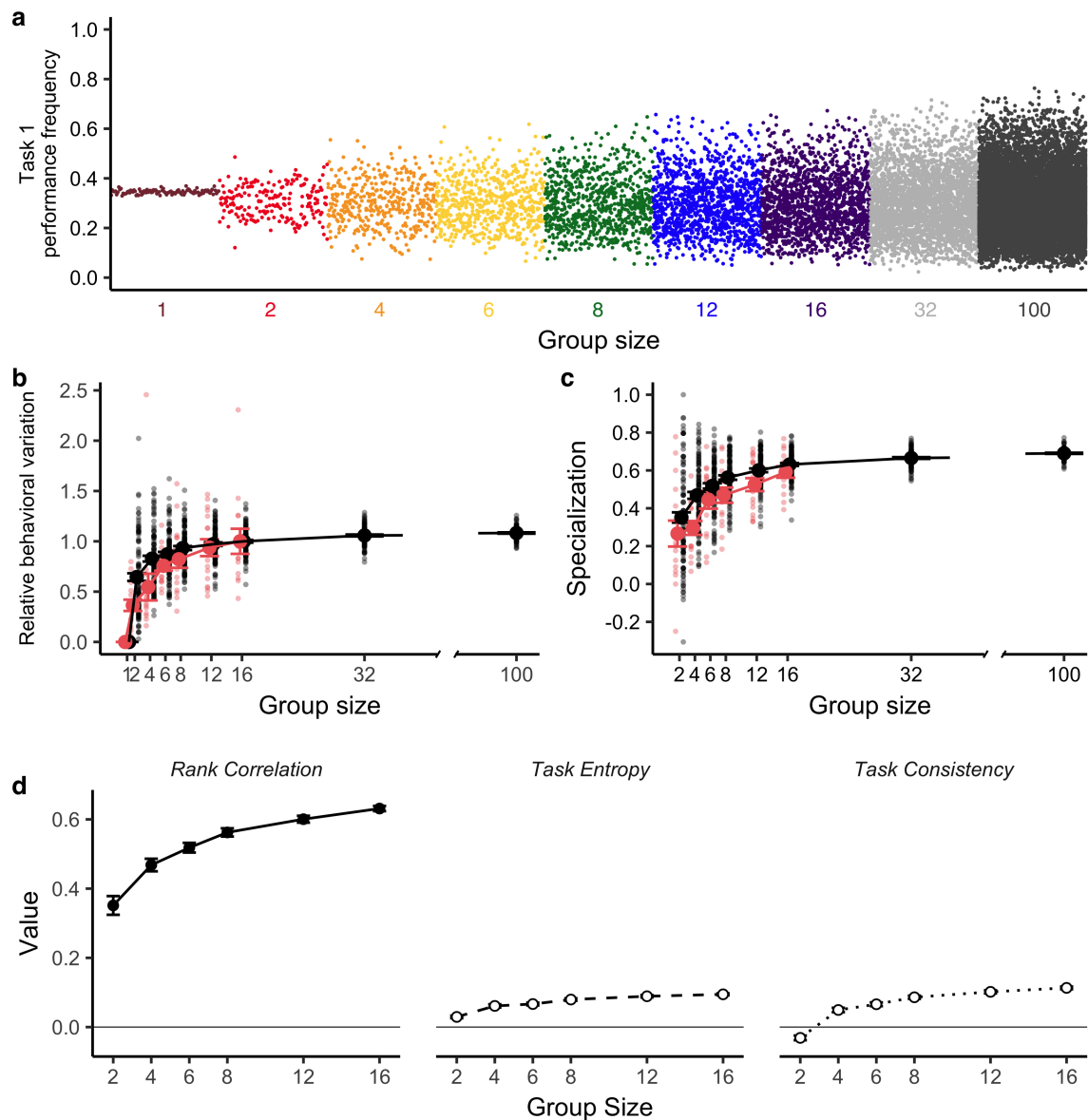
Extended Data Fig. 4 | Effect of the larvae-to-workers ratio on behaviour and brood developmental time. The number of workers was constant at 4, and the number of larvae varied between 4 and 16, so as to obtain larvae-to-workers ratios of 1, 2 or 4. **a**, Mean colony r.m.s.d. increased with the larvae-to-workers ratio (log-transformed r.m.s.d.: $\chi^2 = 5.00$, $P = 0.03$). **b**, Larval time to eclosion was unaffected by the larvae-to-workers ratio (time to eclosion transformed by (time to eclosion)⁵: $\chi^2 = 0.17$, $P = 0.68$). Sample sizes indicate the number of

colonies in which at least one larva reached adulthood. In both panels, box plots represent the median (thick horizontal line); the lower and upper hinges correspond to the first and third quartile, respectively. The upper whiskers extend from the upper hinge to the largest value no further than $1.5 \times$ interquartile range from the hinge; the lower whiskers extend from the lower hinge to the smallest value no further than $1.5 \times$ interquartile range from the hinge.



Extended Data Fig. 5 | Methods for behavioural analyses. a, Resampling scheme. Ninety-five per cent confidence intervals were generated by resampling individual r.m.s.d. values from one colony of size 16 at a time. In the example here, the generation of confidence intervals for behavioural variation (standard deviation of r.m.s.d.) in colonies of size 6 is shown. The same method was used to generate confidence intervals for mean colony behaviour (mean r.m.s.d.). **b–e**, Computing specialization.

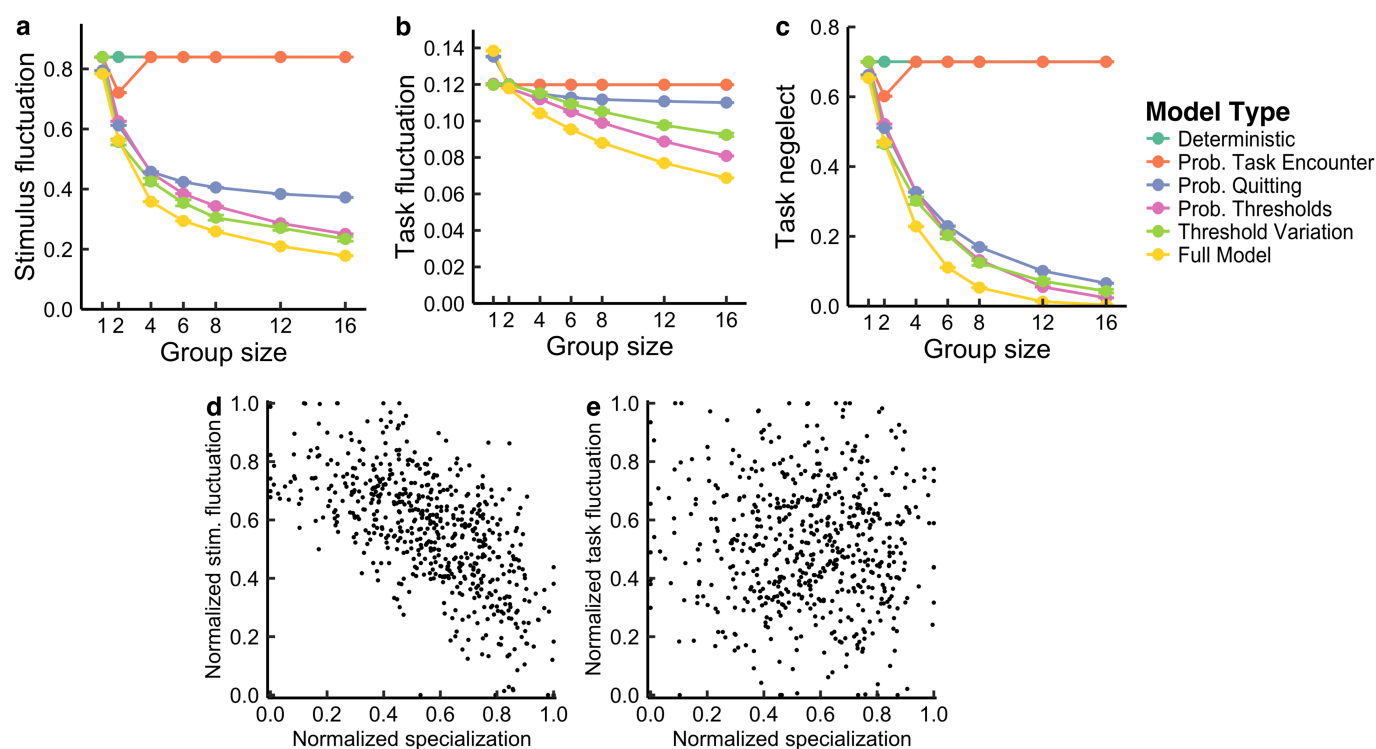
b, Daily individual r.m.s.d. values in one colony of size 8. **c**, Daily individual r.m.s.d. ranks. **d**, Pairwise rank correlation matrix between days of the first brood-care phase. Values highlighted in red indicate rank correlations (Spearman, $n = 16$ ants) between consecutive days, which are averaged to compute short-term behavioural consistency. **e**, Mean r.m.s.d. ranks per brood-care phase used to compute long-term behavioural consistency.



Extended Data Fig. 6 | Behaviour of the fixed threshold model. One hundred replicates were simulated per group size. Parameterization: $m = 2$, $\eta = 7$, $\mu = 10$, $\sigma = 0.1$, $\tau = 0.2$ and $\delta = 0.6$, corresponding to the filled circle symbol in Fig. 3b. **a**, Frequency of task 1 performance (measured across a simulation run) by individual ants at different group sizes; each point represents an ant and ants from the same colony are vertically aligned. **b**, Behavioural variation (standard deviation of individual task performance frequencies) across all 100 replicates for each group size,

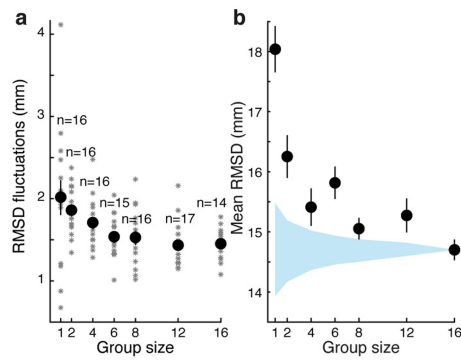
averaged over both tasks and shown relative to group size 16.

c, Specialization in task performance relative to group size. Each point represents one colony, and the line represents the mean value (\pm s.e.m.) across all 100 replicates for each group size. In **b**, **c**, model output is in black and experimental data are in red. **d**, Mean values (\pm s.e.m.) of the rank correlation, task entropy and task consistency metrics across all 100 replicates at each group size.

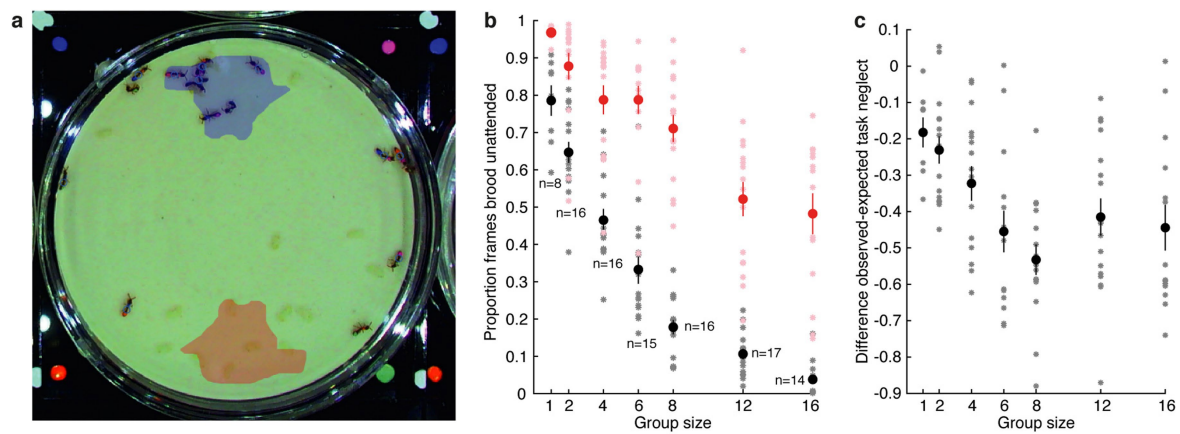


Extended Data Fig. 7 | The effect of stochasticity and specialization on proxies for fitness. One hundred replicates were simulated per group size. Parameter settings for the deterministic model can be found in the Supplementary Methods; departures from deterministic model parameters are as in Extended Data Fig. 6. **a**, Short-term (single time step) stimulus fluctuations averaged across both tasks are shown across group sizes and for all models. **b**, Short-term (single time step) fluctuations in task performance frequency (measured by the proportion of the colony

performing each task), averaged across both tasks, are shown across group sizes and for all models. **c**, Task neglect averaged across both tasks is shown across group sizes and for all models. In **a–c**, points represent the described averages, which have been further averaged (mean \pm s.e.m.) across $n = 100$ replicate colonies of a given size. **d**, **e**, Relationship between specialization and short-term stimulus fluctuations (**d**) or short-term fluctuations in task performance frequency (**e**), in the full model when controlling for group size. Each point represents one simulated colony.

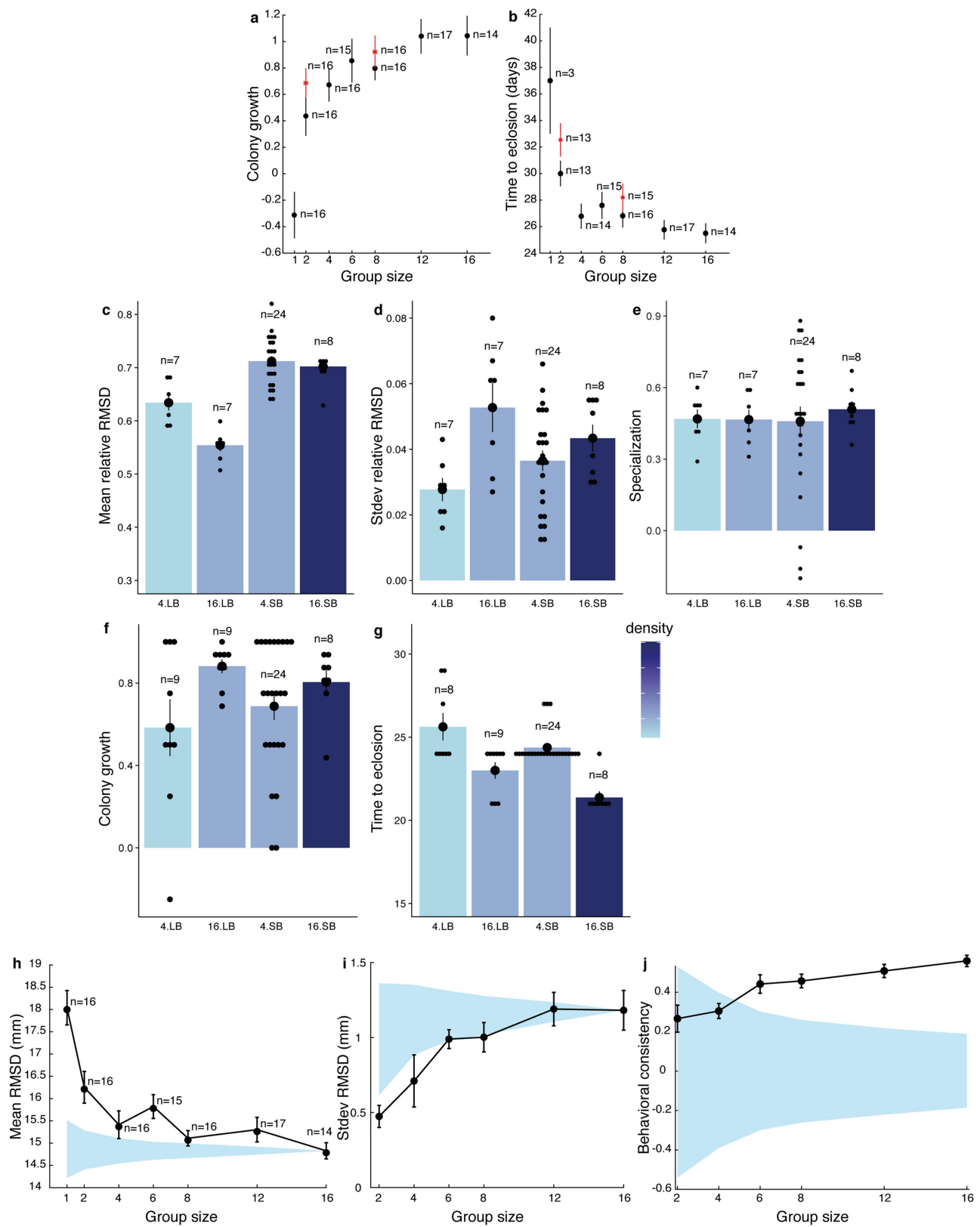


Extended Data Fig. 8 | Behavioural homeostasis increases with group size. **a**, Day-to-day fluctuations in colony mean r.m.s.d. (mean \pm s.e.m.) decrease with group size ($\chi^2 = 21.30$, $P = 3.93 \times 10^{-6}$). Asterisks represent colony-level data. **b**, Mean spatial fidelity increases with group size. Black, colony mean r.m.s.d. as a function of group size (mean \pm s.e.m.). Blue, 95% confidence intervals under the null hypothesis of no group-size effect on individual behaviour, generated by resampling individuals from colonies of size 16 (Extended Data Fig. 5a). Sample sizes are as in **a**. In both panels, data for genotypes A and B are pooled.



Extended Data Fig. 9 | Task neglect. **a**, Manually annotated nest area (blue) and control area (red) generated by rotating the nest area by 180° around the centre of the Petri dish. **b**, Task neglect (mean \pm s.e.m.) decreases with group size. The proportion of frames in which no ant was found near the brood as a function of group size. Black, observed

task neglect. Red, expected task neglect. **c**, Effective task neglect (mean \pm s.e.m.) decreases with group size ($\chi^2 = 13.36$, $P = 2.57 \times 10^{-4}$). The difference between observed and expected task neglect is shown as a function of group size. Sample sizes are as in **b**. In **b**, **c**, data for genotypes A and B are pooled and asterisks represent colonies.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Control experiments. a, b, Paint-marking did not disproportionately affect small colonies. Red asterisks indicate control colonies composed of unmarked ants; otherwise, data are as in Fig. 4b, c. **a,** Growth in colonies of unmarked ants (mean \pm s.e.m.). Colony growth was unaffected by paint-marking ($\chi^2 = 2.71$, $P = 0.10$), the interaction of paint-marking with group size ($\chi^2 = 0.31$, $P = 0.58$) or the interaction of paint-marking with genotype ($\chi^2 = 0.17$, $P = 0.68$). **b,** Larval time-to-eclosion in colonies of unmarked ants (mean \pm s.e.m.). Time to eclosion of larvae was increased by paint-marking of the workers (square-root-transformed time to eclosion: $\chi^2 = 8.98$, $P = 0.003$), but paint-marking did not interact with group size ($\chi^2 = 0.09$, $P = 0.77$) or genotype ($\chi^2 = 0.22$, $P = 0.64$). **c–g,** Effects of density on behaviour and fitness. Colonies consisted of 4 or 16 workers (and a matching number of larvae) in small or large Petri dishes (SB and LB, respectively), corresponding to 3 densities (shades of blue). **c,** Mean spatial fidelity (mean \pm s.e.m.) was affected by group size ($\chi^2 = 6.49$, $P = 0.01$), box size ($\chi^2 = 38.46$, $P = 5.6 \times 10^{-10}$) and density (group size:box size: $\chi^2 = 6.76$, $P = 0.009$). **d,** Behavioural variation (mean \pm s.e.m.) was affected by group size ($\chi^2 = 7.44$, $P = 0.006$) but not by box size ($\chi^2 = 0.08$, $P = 0.77$) or density (group size:box size: $\chi^2 = 3.50$, $P = 0.06$). **e,** Behavioural consistency (mean \pm s.e.m.) was not affected by group size ($\chi^2 = 0.03$, $P = 0.87$), box size ($\chi^2 = 0.22$, $P = 0.64$) or density (group size:box size: $\chi^2 = 0.02$, $P = 0.88$). Behavioural consistency was transformed by (behavioural consistency + 0.21)^{1.5}. **f,** Colony growth (mean \pm s.e.m.) was affected by group size ($\chi^2 = 3.91$, $P = 0.048$), but not

by box size ($\chi^2 = 0.04$, $P = 0.85$) or density (group size:box size: $\chi^2 = 1.00$, $P = 0.32$). Colony growth was transformed by (growth + 0.4)^{1.9}. Thus, the effect of density is small relative to that of group size, and variation in density alone is therefore very unlikely to have confounded our results. **g,** Larval time-to-eclosion (mean \pm s.e.m.) was affected by group size ($\chi^2 = 35.74$, $P = 2.26 \times 10^{-9}$) and box size ($\chi^2 = 10.45$, $P = 0.001$) but not by density (group size:box size: $\chi^2 = 0.67$, $P = 0.41$). Time to eclosion was transformed by (time to eclosion)^{-0.3}. **h–j,** Removing individuals with more than three ovarioles from analyses did not qualitatively affect our results. **h,** Mean spatial fidelity of the colony increases with group size. Black, mean r.m.s.d. (\pm s.e.m.) as a function of group size, after excluding individuals with four or more ovarioles. Blue, 95% confidence interval generated by resampling workers from 16-worker colonies (Extended Data Fig. 5a). **i,** Behavioural variation increases with group size. Black, standard deviation in r.m.s.d. per colony as a function of group size (mean \pm s.e.m.), after excluding individuals with more than three ovarioles. Ninety-five per cent confidence intervals and sample sizes are as in **a**. **j,** Day-to-day rank consistency increases with group size. Black, mean r.m.s.d. rank correlation coefficients over consecutive days in the first brood care phase as a function of group size (mean \pm s.e.m.), after excluding individuals with more than three ovarioles. Blue, 95% confidence intervals generated by randomizing daily ranks in each colony. In **a**, **b**, **h–j**, data for genotypes A and B are pooled.

Global land change from 1982 to 2016

Xiao-Peng Song^{1*}, Matthew C. Hansen¹, Stephen V. Stehman², Peter V. Potapov¹, Alexandra Tyukavina¹, Eric F. Vermote³ & John R. Townshend¹

Land change is a cause and consequence of global environmental change^{1,2}. Changes in land use and land cover considerably alter the Earth's energy balance and biogeochemical cycles, which contributes to climate change and—in turn—affects land surface properties and the provision of ecosystem services^{1–4}. However, quantification of global land change is lacking. Here we analyse 35 years' worth of satellite data and provide a comprehensive record of global land-change dynamics during the period 1982–2016. We show that—contrary to the prevailing view that forest area has declined globally⁵—tree cover has increased by 2.24 million km² (+7.1% relative to the 1982 level). This overall net gain is the result of a net loss in the tropics being outweighed by a net gain in the extratropics. Global bare ground cover has decreased by 1.16 million km² (−3.1%), most notably in agricultural regions in Asia. Of all land changes, 60% are associated with direct human activities and 40% with indirect drivers such as climate change. Land-use change exhibits regional dominance, including tropical deforestation and agricultural expansion, temperate reforestation or afforestation, cropland intensification and urbanization. Consistently across all climate domains, montane systems have gained tree cover and many arid and semi-arid ecosystems have lost vegetation cover. The mapped land changes and the driver attributions reflect a human-dominated Earth system. The dataset we developed may be used to improve the modelling of land-use changes, biogeochemical cycles and vegetation–climate interactions to advance our understanding of global environmental change^{1–4,6}.

Humanity depends on land for food, energy, living space and development. Land-use change—traditionally a local-scale human practice—is increasingly affecting Earth system processes, including the surface energy balance, the carbon cycle, the water cycle and species diversity^{1–4}. Land-use change is estimated to have contributed a quarter of cumulative carbon emissions to the atmosphere since industrialization³. As population and per capita consumption continue to grow, so does demand for food, natural resources and consequent stress to ecosystems.

Because of their synoptic view and recurrent monitoring of the Earth's surface, satellite observations contribute substantially to our current understanding of the global extent and change of land cover and land use. Previous global-scale studies have mainly focused on annual forest cover change (stand-replacement disturbance) for the time period after 2000⁷, or focused on sparse temporal intervals⁸. Long-term gradual changes in undisturbed forests as well as areal changes in cropland, grassland and other non-forested land are less well quantified.

We create an annual, global vegetation continuous fields product⁹ for the time period 1982 to 2016, consisting of tall vegetation (≥ 5 m in height; hereafter referred to as tree canopy (TC)) cover, short vegetation (SV) cover and bare ground (BG) cover, at $0.05^\circ \times 0.05^\circ$ spatial resolution (for details of definitions, see Supplementary Methods). For each year, every land pixel is characterized by its per cent cover of TC, SV and BG, representing the vegetation composition at the time of the local peak growing season. The dataset is produced by combining optical observations from multiple satellite sensors, including the Advanced Very High Resolution Radiometer (AVHRR), the Moderate Resolution

Imaging Spectroradiometer, the Landsat Enhanced Thematic Mapper Plus and various sensors with very high spatial resolution. We use non-parametric trend analysis to detect and quantify changes in tree canopy, short vegetation and bare ground over the full time period at pixel ($0.05^\circ \times 0.05^\circ$), regional and global scales. Observed changes are attributed to direct human activities or indirect drivers on the basis of a global probability sample and interpretation of high-resolution images from Google Earth.

The total area of tree cover increased by 2.24 million km² from 1982 to 2016 (90% confidence interval (CI): 0.93, 3.42 million km²), which represents a +7.1% change relative to 1982 tree cover (Extended Data Table 1). Bare ground area decreased by 1.16 million km² (90% CI: −1.78, −0.34 million km²), which represents a decrease of 3.1% relative to 1982 bare ground cover. The total area of short vegetation cover decreased by 0.88 million km² (90% CI: −2.20, 0.52 million km²), which indicates a decrease of 1.4% relative to 1982 short vegetation cover. A global net gain in tree canopy contradicts current understanding of long-term forest area change; the Food and Agriculture Organization of the United Nations (FAO) reported a net forest loss between 1990 and 2015⁵. However, our gross tree canopy loss estimate (−1.33 million km², −4.2%, Extended Data Table 1) agrees in magnitude with the FAO's estimate of net forest area change (−1.29 million km², −3%), despite differences in the time period covered and definition of forest (the FAO defines 'forest' as tree cover $\geq 10\%$; see details in Supplementary Methods).

The mapped land change (Fig. 1) consists of all changes in land cover and land use induced by natural or anthropogenic drivers. Land change themes are also inherently linked in the tree cover–short vegetation–bare ground nexus. For example, deforestation for agricultural expansion is often manifested as tree canopy loss and short vegetation gain, whereas land degradation may simultaneously result in short vegetation loss and bare ground gain. Pairs of changes in TC (ΔTC), SV (ΔSV) and BG (ΔBG) show strong coupling and symmetry in change direction but vary substantially over space (Fig. 1b and Extended Data Fig. 1). That is, the globally dominant, coupled land changes are ΔTC co-located with ΔSV and ΔSV co-located with ΔBG .

The overall net gain in tree canopy is a result of a net loss in the tropics being outweighed by a net gain in the subtropical, temperate and boreal climate zones (Extended Data Table 2). A latitudinal north (gain)–south (loss) contrast in tree cover change is evident (Fig. 2a). Conversely, for short vegetation tropical net gain is exceeded by extratropical net loss. The latitudinal profile of ΔSV largely mirrors that of ΔTC , most obviously in the northern mid-to-high latitudes ($45^\circ N$ – $75^\circ N$) and low latitudes ($30^\circ S$ – $10^\circ N$) (Fig. 2b). For bare ground, subtropical net gain partially offsets losses in all other climate domains. In the northern low-to-mid latitudes ($10^\circ N$ – $45^\circ N$), the profile of bare ground loss (Fig. 2c) closely corresponds to that of short vegetation gain (Fig. 2b).

Changes were unevenly distributed across biomes (Fig. 3, Extended Data Fig. 2 and Extended Data Table 2). The largest area of net tree canopy loss occurred in the tropical dry forest biome (−95,000 km², −8%) (Extended Data Fig. 2a), closely followed by tropical moist deciduous forest (−84,000 km², −2%) (Fig. 3c) (all per cent net changes

¹Department of Geographical Sciences, University of Maryland, College Park, MD, USA. ²College of Environmental Science and Forestry, State University of New York, Syracuse, NY, USA. ³NASA Goddard Space Flight Center, Greenbelt, MD, USA. *e-mail: xpsong@umd.edu

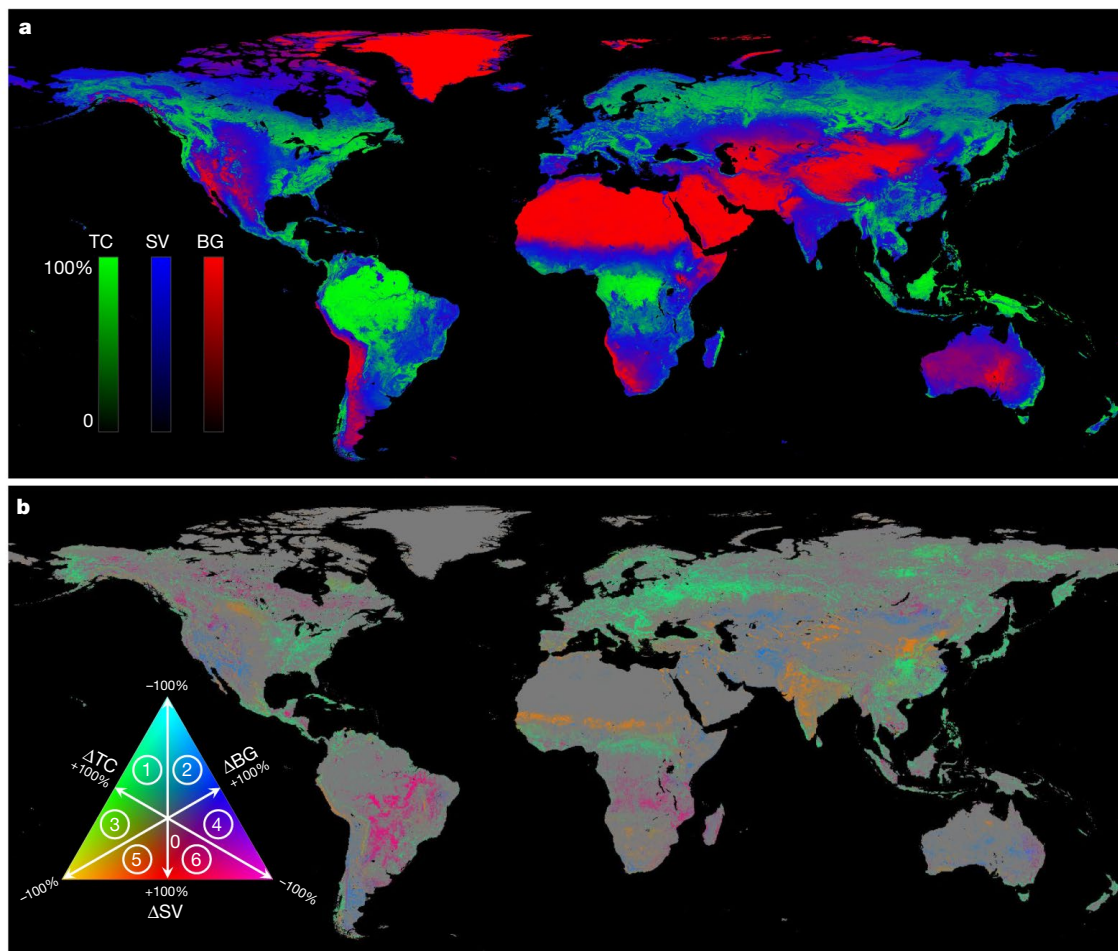


Fig. 1 | A satellite-based record of global TC, SV and BG cover from 1982 to 2016. a, Mean annual estimates. b, Long-term change estimates. Both mean and change estimates are expressed as per cent of pixel area at $0.05^\circ \times 0.05^\circ$ spatial resolution. Pixels showing a statistically significant trend ($n = 35$, two-sided Mann–Kendall test, $P < 0.05$) in either TC, SV or

BG are depicted on the change map. Circled numbers in the colour legend denote dominant change directions: 1, TC gain with SV loss; 2, BG gain with SV loss; 3, TC gain with BG loss; 4, BG gain with TC loss; 5, SV gain with BG loss; and 6, SV gain with TC loss.

are expressed relative to the benchmark of the area of the cover class in 1982). Tree canopy in major forest biomes outside the tropics has increased over the past 35 years: temperate continental forest has experienced the largest gain ($+726,000 \text{ km}^2$, $+33\%$) (Fig. 3d), which is comparable to the next two biomes—boreal coniferous forest ($+463,000 \text{ km}^2$, $+12\%$) and subtropical humid forest ($+280,000 \text{ km}^2$, $+18\%$)—combined (Extended Data Fig. 2e, m).

Short vegetation loss mirrored tree cover gain dynamics, but with smaller magnitudes: temperate continental forest ($-610,000 \text{ km}^2$, -14%), boreal coniferous forest ($-430,000 \text{ km}^2$, -10%) and subtropical humid forest ($-249,000 \text{ km}^2$, -9%). By contrast, tropical forest biomes all gained short vegetation, with tropical shrubland experiencing the largest areal increase ($+417,000 \text{ km}^2$, $+10\%$) (Fig. 3e), twice the amount of short vegetation gain in tropical dry forest ($+246,000 \text{ km}^2$, $+5\%$). Tropical shrubland also experienced the largest bare ground loss ($-408,000 \text{ km}^2$, -10%). Subtropical desert—the second largest dryland biome on Earth—had the largest gain in bare ground ($+154,000 \text{ km}^2$, $+4\%$) (Fig. 3f), followed by subtropical steppe ($+107,000 \text{ km}^2$, $+5\%$) (Extended Data Fig. 2h).

Consistently across all climate domains, mountain systems experienced net bare ground loss, net short vegetation loss and net tree canopy gain (Extended Data Fig. 2c, f, i, n and Extended Data Table 2). In the high-latitude boreal tundra woodland and the polar ecozone (Extended Data Fig. 2o, p), bare ground decreased and tree canopy increased in both biomes, whereas short vegetation decreased in tundra woodland but increased in the polar ecozone.

Based on the data from the global probability sample, an estimated 60% of all changes were associated with direct human land-use activities and 40% with indirect drivers such as climate change (Extended Data Figs. 3, 4; see Supplementary Methods). Direct human impact varied from 36% for bare ground gain to 70% for tree canopy loss. At the

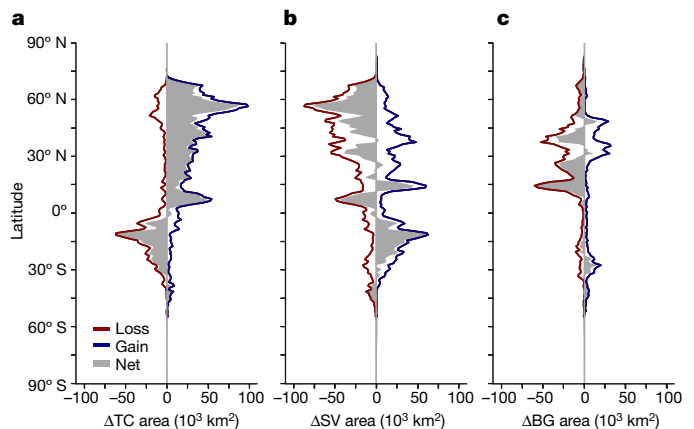


Fig. 2 | Latitudinal profiles of change in land cover from 1982 to 2016. a, Tree canopy cover change (ΔTC). b, Short vegetation cover change (ΔSV). c, Bare ground cover change (ΔBG). Area statistics were calculated for every 1° of latitude.

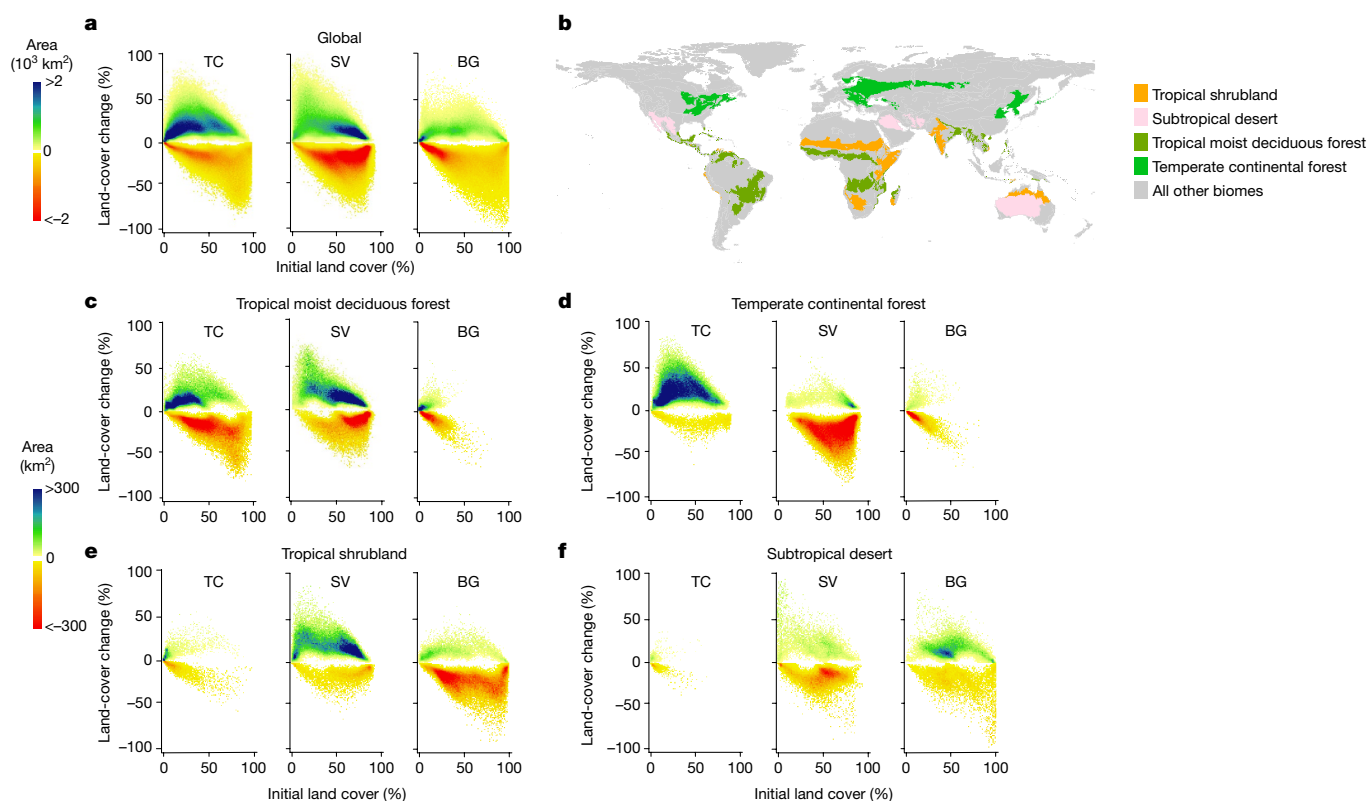


Fig. 3 | Intensity plots of gross area of loss and gain in TC, SV and BG cover during 1982–2016. **a**, Global-scale plots (top colour bar). Initial land cover (x axis) is defined as mean value of the first five years, 1982–1986. To create these plots, for each cover class, per cent change layer (Fig. 1b) and initial cover layer are used to construct a 2D histogram with bin size of 1% for both axes. Then, total change area in each bin is calculated and plotted. Data points located towards the lower-right corner of the TC plot are more likely to be deforestation (that is, points with large initial tree cover and large reduction in tree cover). The concentrated blue region of the SV plots reflects cropland intensification. The green belt on the BG plot suggests that vegetation loss occurred across the entire range

of BG coverage. The dominance of TC gain over TC loss, SV loss over SV gain and BG loss over BG gain are also clearly revealed. **b**, Geographical distribution of four highlighted biomes with largest gross areal changes. Biome distribution from a previous publication³⁰, reproduced with permission. **c**, Largest gross TC loss and SV gain. **d**, Largest gross TC gain and SV loss. **e**, Largest gross BG loss. **f**, Largest gross BG gain. The bottom colour bar is consistent across biomes (**c**–**f**) and cover types. Long-term gross dynamics of TC, SV and BG changes vary considerably between biomes. See Extended Data Fig. 2 for other biomes and Extended Data Table 2 for change area estimates.

continental scale, land-use activities account for the majority of observed land changes in Europe (86%), South America (66%), Asia (62%) and Africa (50%), but have a smaller role in North America (47%) and Oceania (35%). The specific land-change drivers are diverse, multi-scale and interactive¹, as discussed in detail below. However, changes collectively induced by the various drivers at the global scale appear to have been gradual over time (Extended Data Fig. 5).

Expansion of the agricultural frontier is the primary driver of deforestation in the tropics¹⁰. The three countries with the largest area of net tree cover loss during 1982–2016 are all located in South America: Brazil (−385,000 km², −8%), Argentina (−113,000 km², −25%) and Paraguay (−79,000 km², −34%) (Supplementary Table 1). The ‘arc of deforestation’ along the southeastern edge of the Amazon has been well-documented^{7,10}. Clearing of natural vegetation for export-oriented industrial agriculture also prevailed in the Cerrado (Fig. 4a) and the Gran Chaco (Fig. 4b). Spatially clustered hotspots of deforestation are also found in Queensland, Australia, and in Southeast Asia—including Myanmar, Vietnam, Cambodia and Indonesia—diminishing the already scarce primary forests of the region¹¹. In sub-Saharan Africa, tree cover loss was pervasive across the Congolian rainforests and the Miombo woodlands (Fig. 4c), historically related to smallholder agriculture and increasingly to commodity crop cultivation¹². Forests in boreal Canada, eastern Alaska and central Siberia exhibited large patches of tree canopy loss and short vegetation gain, similar to the tropics (Fig. 1b). However, these are the result of persistent disturbances from wildfires and subsequent recovery of natural vegetation¹³.

Discernible effects of climate change on vegetation change are also revealed at regional scales. In the western United States (Fig. 4d), forests are suffering from increasing stress from insects, wildfires, heat and droughts due to regional warming¹⁴. But in the temperature-limited Arctic, warming is facilitating woody vegetation growth in northeastern Siberia, western Alaska and northern Quebec¹⁵ (Fig. 4e). Land-use activities are rare in these boreal tundra and polar ecosystems, contributing less than 1% to observed land changes (Extended Data Fig. 3e). In water-limited savannahs in Central and West Africa (Fig. 4f), forest expansion and woody encroachment—observed both from space and in the field¹⁶—are probably driven by increases in precipitation and atmospheric carbon dioxide¹⁷. Extreme high-rainfall anomalies also contributed to the greening of the Sahel¹⁷ (Fig. 4f). Altitudinal biome shift is also expected in a changing climate. Global treeline positions have been advancing since AD 1900 as a result of climate warming¹⁸. The aforementioned bare ground loss, short vegetation loss and tree canopy gain in global mountain systems further suggest that an enduring transformation is occurring with regard to the distribution, structure and composition of montane vegetation.

Political, social and economic factors can influence vegetation in conjunction with climate drivers. Tree canopy in Europe, including European Russia, has increased by 35%—the greatest gain among all continents (Extended Data Table 1). Spatially contiguous hotspots of tree canopy gain were found in European Russia and Carpathian montane forests (Fig. 4g). Natural afforestation on abandoned agricultural land has been a common process in Eastern Europe after the collapse

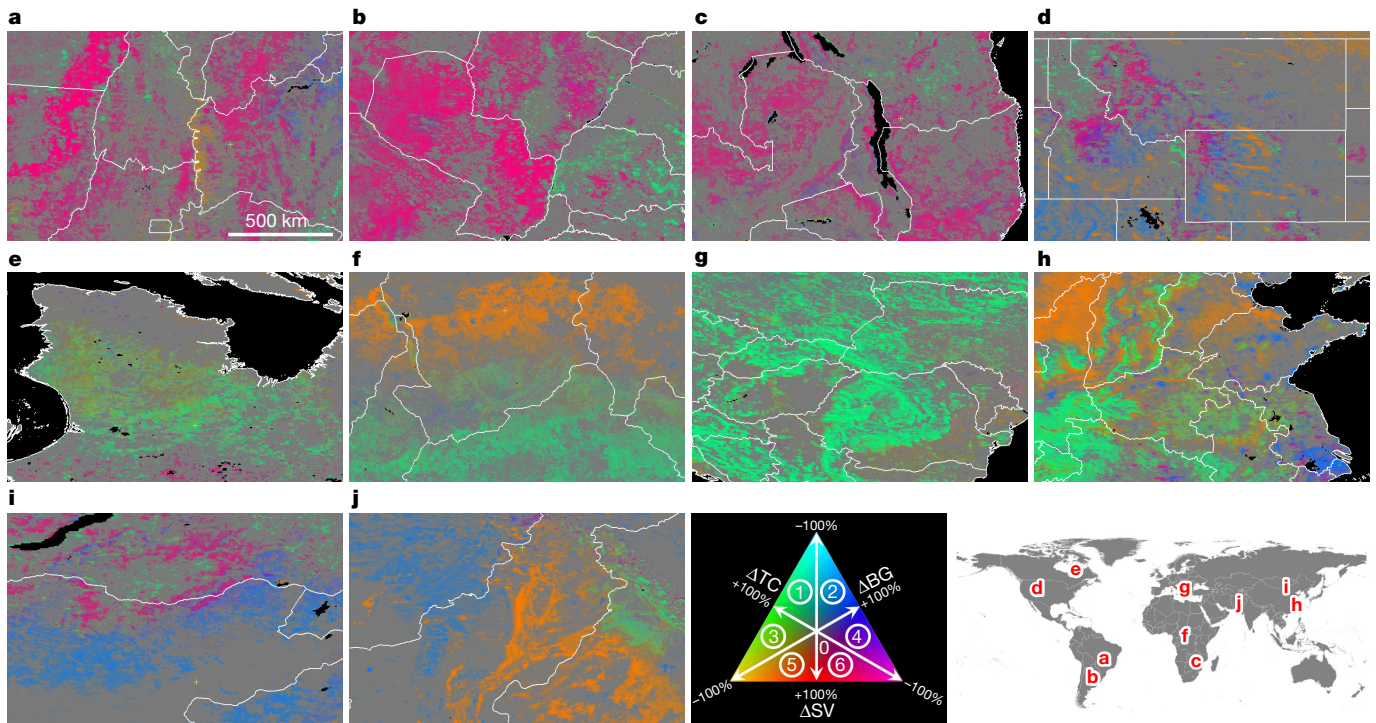


Fig. 4 | Regional subsets of changes in TC, SV and BG cover. As in Fig. 1b, pixels showing a statistically significant trend ($n = 35$, two-sided Mann–Kendall test, $P < 0.05$) in TC, SV or BG are depicted on the change map. **a**, Cerrado ecoregion in Brazil, centred at 11.4° S, 46.5° W. **b**, Gran Chaco ecoregion in Bolivia, Argentina and Paraguay, centred at 22.5° S, 55.7° W. **c**, Miombo woodlands in southeast Africa, centred at 12.4° S, 33.9° E. **d**, Western United States, centred at 44.5° N, 110.0° W. **e**, Quebec, Canada,

centred at 57.9° N, 71.6° W. **f**, Central Africa, centred at 10.4° N, 19.4° E. **g**, Eastern Europe, centred at 46.1° N, 20.3° E. **h**, Eastern China, centred at 35.0° N, 115.1° E. **i**, Eastern Mongolia, centred at 48.7° N, 111.0° E. **j**, Afghanistan and Pakistan, centred at 30.7° N, 70.6° E. Circled numbers in the colour legend denote dominant change directions: 1, TC gain with SV loss; 2, BG gain with SV loss; 3, TC gain with BG loss; 4, BG gain with TC loss; 5, SV gain with BG loss; and 6, SV gain with TC loss.

of the Soviet Union¹⁹. Our satellite record confirms the effectiveness of China's large-scale reforestation and afforestation programs, particularly in the Loess Plateau and the Qin Ling–Daba Mountains²⁰ (Fig. 4h). An increasing area of plantations in southeastern China has also led to tree canopy gain (+34%) in China. Tree canopy also increased in the United States (+15%), mostly in the eastern United States (Fig. 1b). Unlike declining forest cover in the western United States (Fig. 4d), southeastern forests are recovering from historical disturbances or are under intensive forestry management²¹.

The world's arid and semi-arid drylands exhibited large areas of decrease in short vegetation and large areas of increase in bare ground, indicating long-term land degradation. Hotspots of vegetation loss include the southwestern United States, southern Argentina, Kazakhstan, Mongolia (Fig. 4i), Inner Mongolia, China, Afghanistan (Fig. 4j) and large areas of Australia. The decrease in short vegetation cover in eastern Australia is probably the consequence of the long-term precipitation decline in the local growing season²². Rising surface temperatures, a reduction in rainfall, and overgrazing caused extensive grassland deterioration in the Mongolian steppe²³. A nationwide ground survey in the United States revealed degradation of soils and vegetation combined with an increased dominance of invasive species in the southwest²⁴.

Human activities undoubtedly have a dominant role in agricultural and urban landscapes, where lands have been continually modified throughout human history. India and China had the largest bare ground loss among all countries (India, $-270,000$ km², -34% ; China, $-250,000$ km², -7%). India also ranked second in short vegetation gain ($+195,000$ km², $+9\%$), after Brazil ($+396,000$ km², $+12\%$). While the short vegetation gain in Brazil is mainly due to the expansion of agricultural frontiers into natural ecosystems, short vegetation gain in India is primarily due to intensification of existing agricultural lands—a continuation of the 'Green Revolution'²⁵. Some of the observed bare ground

gain can be attributed to resource extraction and urban sprawl, most notably in eastern China (Fig. 4h). However, at the global scale, the growth of urban areas accounts for a small fraction of all land changes²⁶.

Previous studies have found a greening Earth on the basis of trends in satellite-based vegetation properties (for example, leaf area index) and have linked this greening trend to a number of climatic and ecological factors^{20,27–29}. A recent study²⁹ using ecosystem models attributed 70% of the observed increase in the global leaf area index to the CO₂ fertilization effect and 4% to land-use change. Our finding that global bare ground cover has decreased over the past 35 years suggests a net increase in vegetation cover and is thus consistent with the greening trend. However, our results differ from previous studies by quantifying the prominent role of land use in global vegetation change. Using a global probability-based sample, we attribute 60% of observed land changes to land-use activities (Extended Data Fig. 3). Our empirical approach is based on observations of high-resolution satellite data (Extended Data Fig. 4), avoiding the challenges of modelling the underlying drivers of land change¹. Additionally, our TC–SV–BG land-cover product is thematically more advanced than vegetation indices in characterizing land surface change. For example, differentiating long-term changes in tree cover from other vegetation can facilitate an improved understanding of global fluxes of water, carbon and energy⁹. Our study provides observational evidence of increasing tree cover in northern continents, which may constitute the missing carbon sink³. By contrast, tropical tree cover loss is associated with higher biomass forests and is responsible for carbon emissions from deforestation^{3,5}. These satellite-based trends are substantiated through the uncertainty analyses (Extended Data Fig. 6; see Supplementary Methods), with the caveat that the long-term field data that would be ideal for verifying historical land-cover change are not available.

The results of this study reflect a human-dominated Earth system. Direct human action on landscapes is found over large areas on every

continent, from intensification and extensification of agriculture to increases in forestry and urban land uses, with implications for the maintenance of ecosystem services². However, human-induced climate change has been documented as an indirect cause of many of the quantified large-scale regional change dynamics, including woody encroachment in Arctic and montane systems and vegetation loss in semi-arid ecoregions^{15,17,18,22,23,29}. Continuing land-use change and the increasing role of climate change in modifying land cover warrants continued monitoring of the Earth's land surface from space.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The AVHRR vegetation continuous fields products that we generated will be distributed through Land Processes Distributed Active Archive Center (LP DAAC, https://lpdaac.usgs.gov/dataset_discovery/measures/measures_products_table/vcf5kyr_v001). Vegetation continuous fields change and uncertainty layers are also provided at <https://glad.umd.edu/dataset/long-term-global-land-change> for download. All other data are available from the corresponding author upon reasonable request.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0411-9>.

Received: 13 September 2017; Accepted: 4 July 2018;

Published online 8 August 2018.

- Turner, B. L. II, Lambin, E. F. & Reenberg, A. The emergence of land change science for global environmental change and sustainability. *Proc. Natl Acad. Sci. USA* **104**, 20666–20671 (2007).
- Foley, J. A. et al. Global consequences of land use. *Science* **309**, 570–574 (2005).
- Le Quéré, C. et al. Global carbon budget 2016. *Earth Syst. Sci. Data* **8**, 605–649 (2016).
- Alkama, R. & Cescatti, A. Biophysical climate impacts of recent changes in global forest cover. *Science* **351**, 600–604 (2016).
- FAO. *Global Forest Resources Assessment 2015* (UN Food and Agriculture Organization, Rome, 2015).
- Bonan, G. B. & Doney, S. C. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science* **359**, eaam8328 (2018).
- Hansen, M. C. et al. High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).
- Feng, M. et al. Earth science data records of global forest cover and change: assessment of accuracy in 1990, 2000, and 2005 epochs. *Remote Sens. Environ.* **184**, 73–85 (2016).
- DeFries, R. S. et al. Mapping the land surface for global atmosphere–biosphere models: toward continuous distributions of vegetation's functional properties. *J. Geophys. Res.* **100**, 20867–20882 (1995).
- Gibbs, H. K. et al. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proc. Natl Acad. Sci. USA* **107**, 16732–16737 (2010).
- Margono, B. A., Potapov, P. V., Turubanova, S., Stolle, F. & Hansen, M. C. Primary forest cover loss in Indonesia over 2000–2012. *Nat. Clim. Change* **4**, 730–735 (2014).
- Ordway, E. M., Asner, G. P. & Lambin, E. F. Deforestation risk due to commodity crop expansion in sub-Saharan Africa. *Environ. Res. Lett.* **12**, 044015 (2017).
- Hicke, J. A. et al. Postfire response of North American boreal forest net primary productivity analyzed with satellite observations. *Glob. Change Biol.* **9**, 1145–1157 (2003).
- van Mantgem, P. J. et al. Widespread increase of tree mortality rates in the western United States. *Science* **323**, 521–524 (2009).
- McManus, K. M. et al. Satellite-based evidence for shrub and graminoid tundra expansion in northern Quebec from 1986 to 2010. *Glob. Change Biol.* **18**, 2313–2323 (2012).
- Mitchard, E. T. & Flintrop, C. M. Woody encroachment and forest degradation in sub-Saharan Africa's woodlands and savannas 1982–2006. *Phil. Trans. R. Soc. Lond. B* **368**, 20120406 (2013).
- Brandt, M. et al. Human population growth offsets climate-driven increase in woody vegetation in sub-Saharan Africa. *Nat. Ecol. Evol.* **1**, 0081 (2017).
- Harsch, M. A., Hulme, P. E., McGlone, M. S. & Duncan, R. P. Are treelines advancing? A global meta-analysis of treeline response to climate warming. *Ecol. Lett.* **12**, 1040–1049 (2009).
- Potapov, P. V. et al. Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. *Remote Sens. Environ.* **159**, 28–43 (2015).
- Piao, S. et al. Detection and attribution of vegetation greening trend in China over the last 30 years. *Glob. Change Biol.* **21**, 1601–1609 (2015).
- Birdsey, R., Pregitzer, K. & Lucier, A. Forest carbon management in the United States: 1600–2100. *J. Environ. Qual.* **35**, 1461–1469 (2006).
- Donohue, R. J., McVicar, T. R. & Roderick, M. L. Climate-related trends in Australian vegetation cover as inferred from satellite observations, 1981–2006. *Glob. Change Biol.* **15**, 1025–1039 (2009).
- Liu, Y. Y. et al. Changing climate and overgrazing are decimating Mongolian steppes. *PLoS ONE* **8**, e57599 (2013).
- Herrick, J. E. et al. National ecosystem assessments supported by scientific and local knowledge. *Front. Ecol. Environ.* **8**, 403–408 (2010).
- Evenson, R. E. & Gollin, D. Assessing the impact of the green revolution, 1960 to 2000. *Science* **300**, 758–762 (2003).
- Ying, Q. et al. Global bare ground gain from 2000 to 2012 using Landsat imagery. *Remote Sens. Environ.* **194**, 161–176 (2017).
- Myneni, R. B., Keeling, C. D., Tucker, C. J., Asrar, G. & Nemani, R. R. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature* **386**, 698–702 (1997).
- Forkel, M. et al. Codominant water control on global interannual variability and trends in land surface phenology and greenness. *Glob. Change Biol.* **21**, 3414–3435 (2015).
- Zhu, Z. et al. Greening of the Earth and its drivers. *Nat. Clim. Change* **6**, 791–795 (2016).
- FAO. *Global Ecological Zoning for the Global Forest Resources Assessment 2000* <http://www.fao.org/docrep/006/ad652e/ad652e00.htm> (UN Food and Agriculture Organization, Rome, 2001).

Acknowledgements This study was funded by the NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program (NNX13AJ35A), Gordon and Betty Moore Foundation (5131), Norwegian Climate and Forests Initiative through the World Resources Institute's Global Forest Watch project, DOB Ecology through the World Resources Institute's Global Restoration Initiative, the NASA Land-Cover and Land-Use Change (LCLUC) Program (NNX15AK65G), and the NASA Carbon Monitoring Systems Program (NNX13AP48G). We thank T. Loveland, B. Pengra and P. Olofsson for making their tree cover validation data available, C. Dimiceli for assistance with vegetation continuous field development and Z. Song for assistance with AVHRR calibration.

Reviewer information Nature thanks M. Forkel, L. Zhou and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions X.-P.S. and M.C.H. designed the study; X.-P.S. carried out data analysis; X.-P.S., M.C.H. and S.V.S. wrote the article with contributions from all authors.

Competing interests The authors declare no competing interests.

Additional information

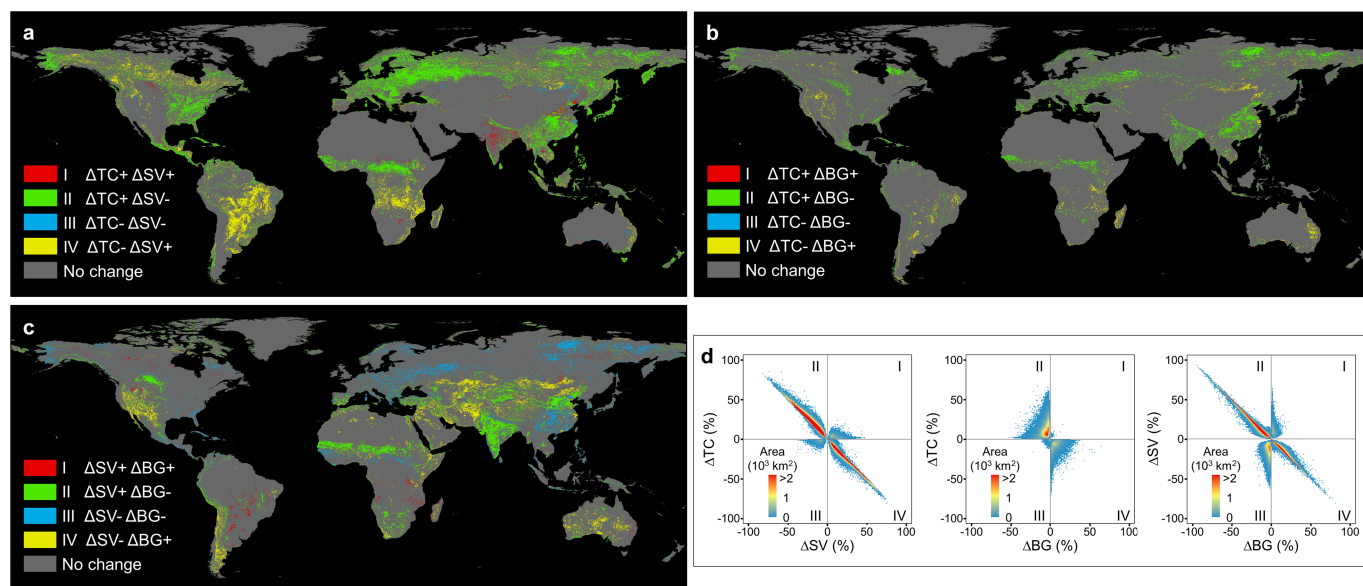
Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0411-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0411-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

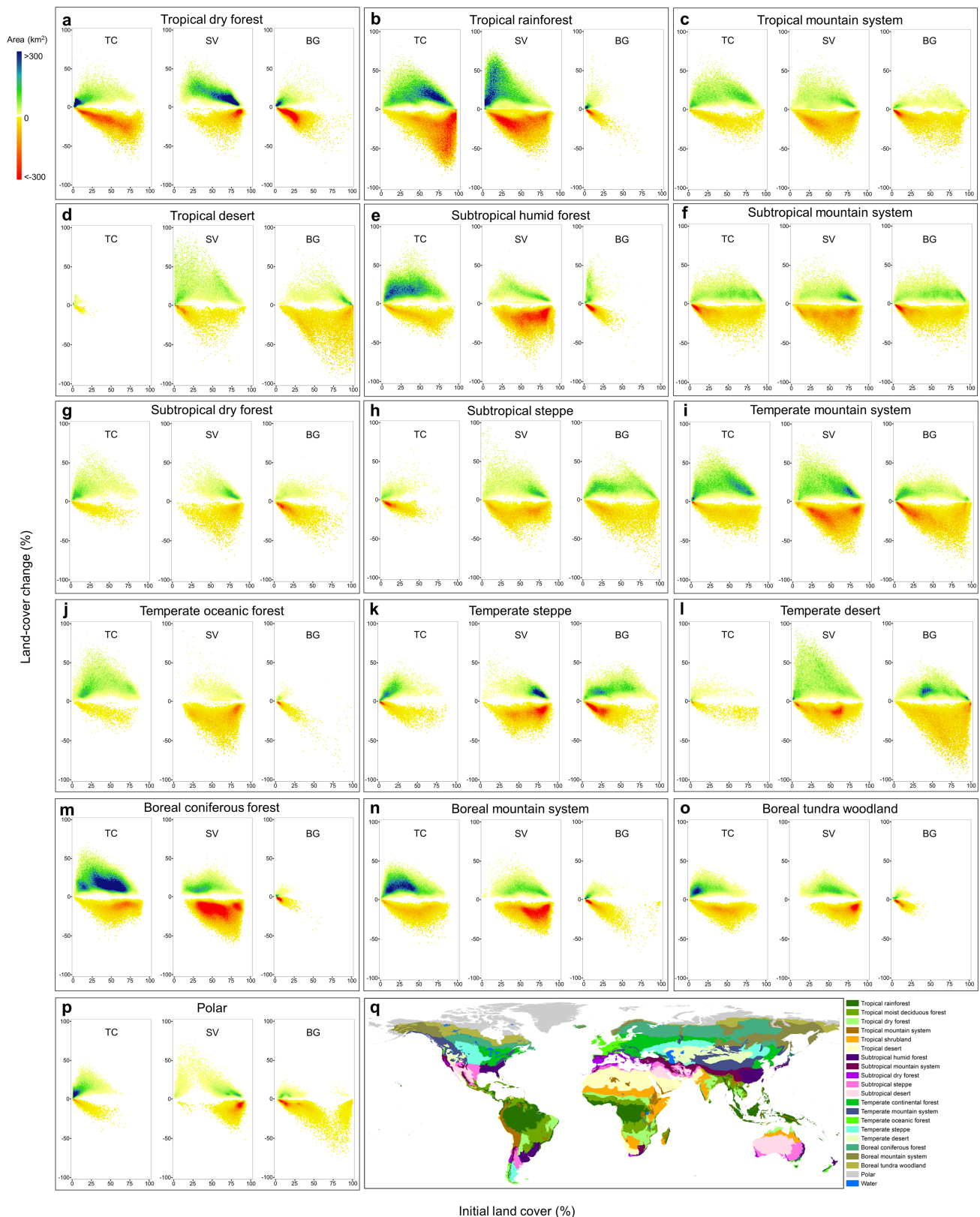
Correspondence and requests for materials should be addressed to X.-P.S.

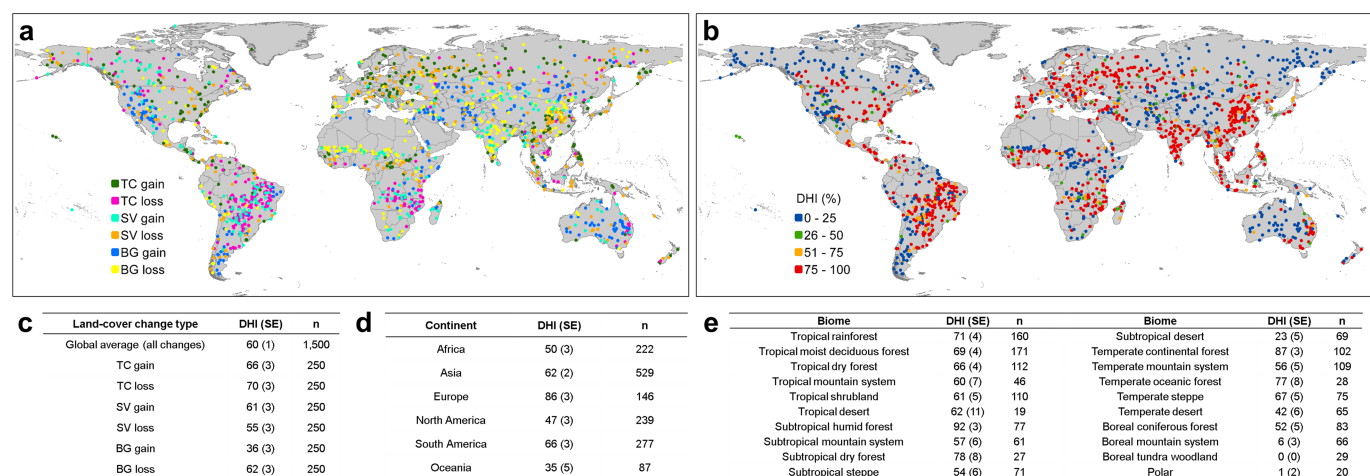
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Satellite-derived, long-term (1982–2016) changes in land cover show strong coupling and symmetry in change detection. **a**, Global map of co-located ΔTC and ΔSV . Pixels showing a statistically significant trend ($n = 35$ years, two-sided Mann–Kendall test, $P < 0.05$) in both TC and SV are depicted on the map. **b**, Global map of co-located ΔTC and ΔBG . **c**, Global map of co-located ΔSV and ΔBG . **d**, From left to

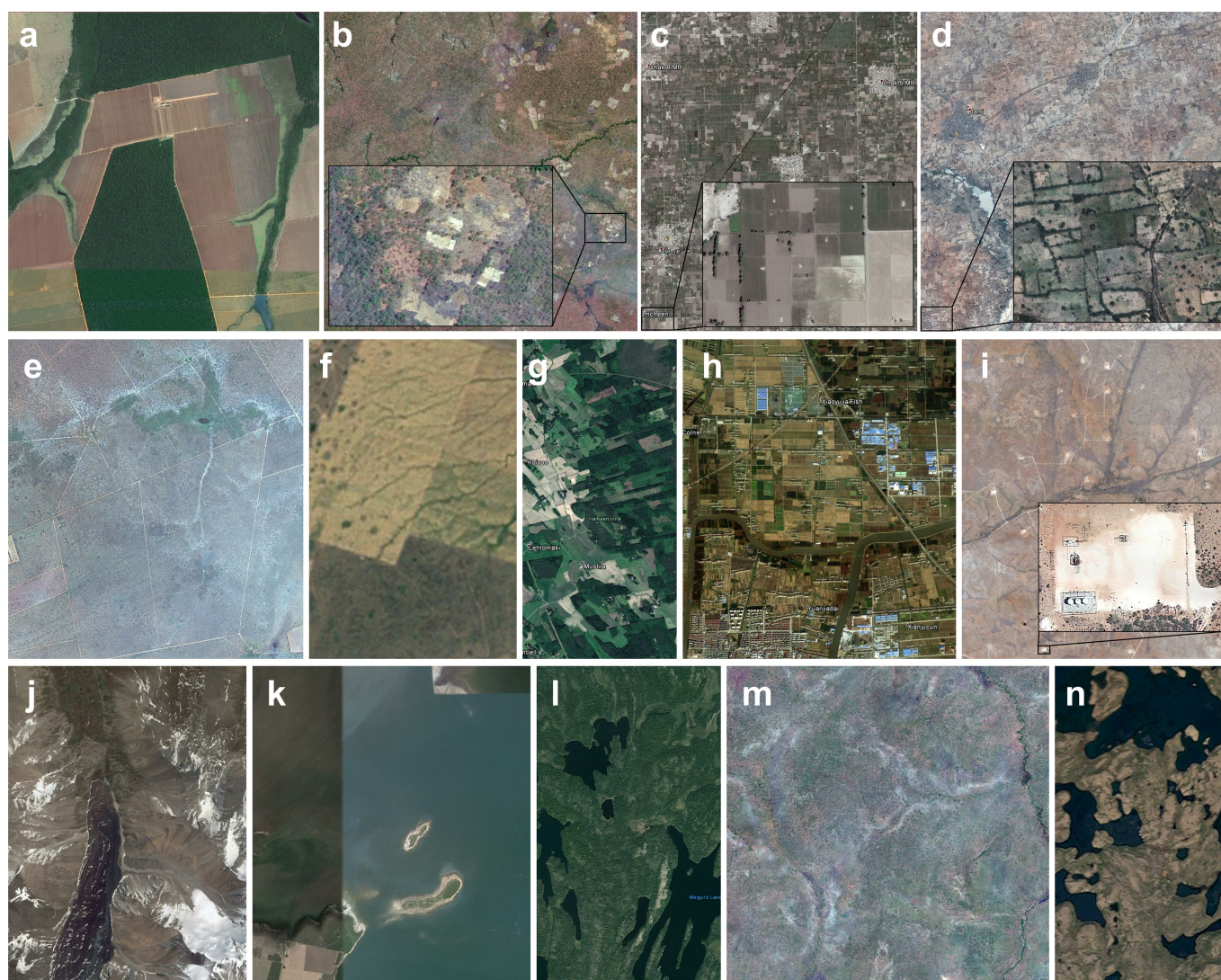
right, intensity plot of change area for ΔTC versus ΔSV , ΔTC versus ΔBG and ΔSV versus ΔBG , corresponding to **a**, **b** and **c**, respectively. To create these intensity plots, paired per cent change layers (Fig. 1b) are used to construct a 2D histogram with bin size of 1% for both axes. Then, the total change area in each bin is calculated and plotted.





Extended Data Fig. 3 | Attributing direct human impact versus indirect drivers to detected changes in land cover. Indirect drivers include both natural drivers and human-induced climate change. **a**, Spatial distribution of the probability sample used for the attribution estimates ($n = 1,500$). **b**, Direct human impact (DHI) of each sample unit interpreted using a time-series of high-resolution images in Google Earth. **c**, Estimated

DHI as a per cent of all change area at the global scale. Global average is calculated by weighting the human impact of each type by each respective global total area provided in Extended Data Table 1. The standard error (SE) for the estimated per cent of DHI is provided in the parentheses. **d**, **e**, Estimated DHI at the continental and biome scales. See Extended Data Fig. 4 for some representative sample examples.

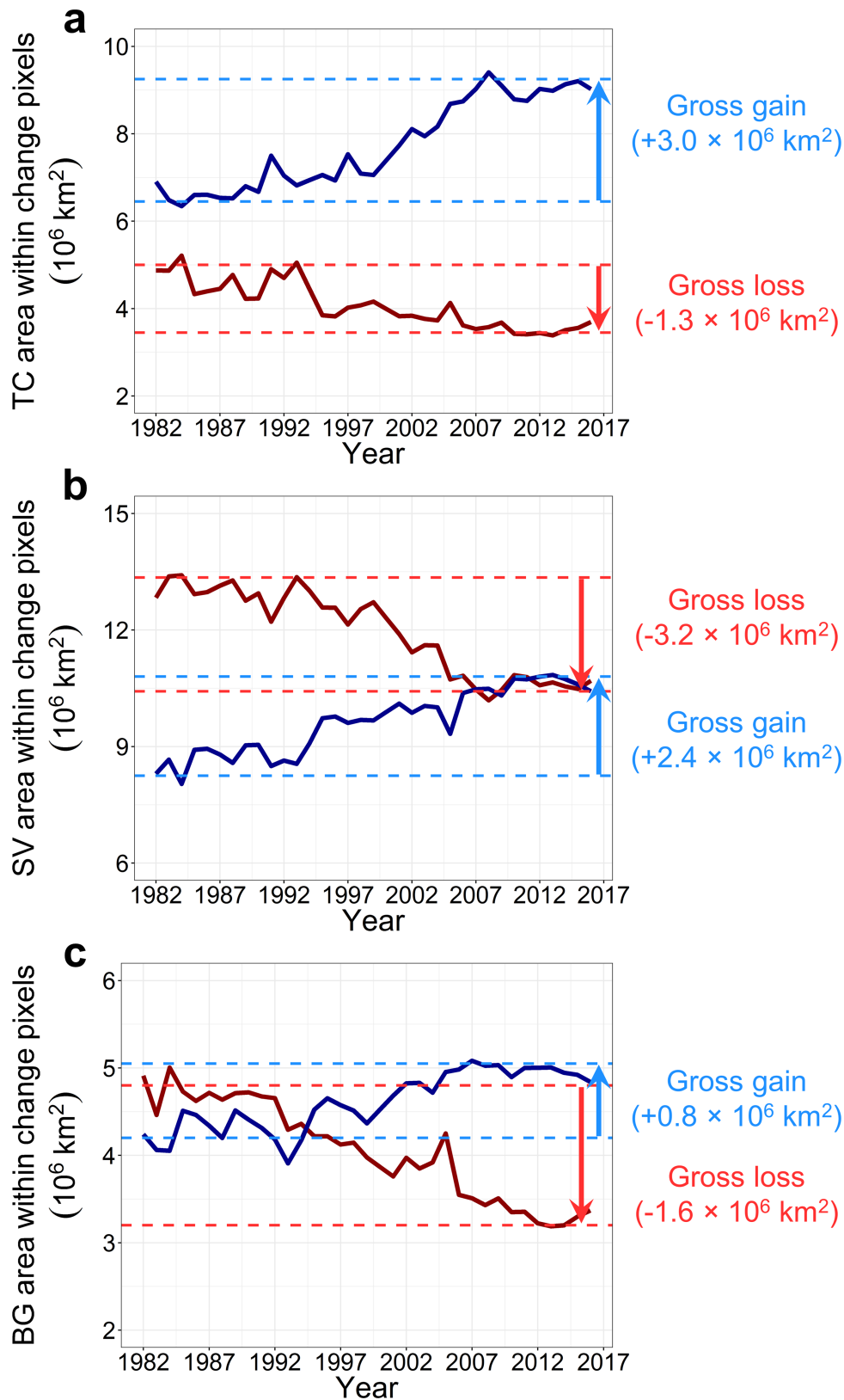


Extended Data Fig. 4 | Selected sample examples for driver attribution.

Screenshots are taken from Google Earth. Each panel is $0.05^\circ \times 0.05^\circ$ in size, corresponding to one AVHRR pixel. **a**, Deforestation for industrial agriculture expansion in Mato Grosso, Brazil (11.275° S, 52.125° W). **b**, Expanding shifting agriculture in northern Zambia (11.625° S, 28.625° E). **c**, Intensification of small-holder agriculture in Punjab, Pakistan (30.025° N, 71.675° E). **d**, Short vegetation gain in low-intensity agricultural lands in northern Nigeria (12.825° N, 7.825° E). **e**, Short vegetation increase due to effective fire suppression in pasture lands in Omaheke, Namibia³¹ (22.175° S, 18.925° E). **f**, Managed pasture lands in western Kazakhstan (49.475° N, 47.725° E). **g**, Forestry in southern Finland (61.075° N, 24.475° E). **h**, Urbanization in Shanghai,

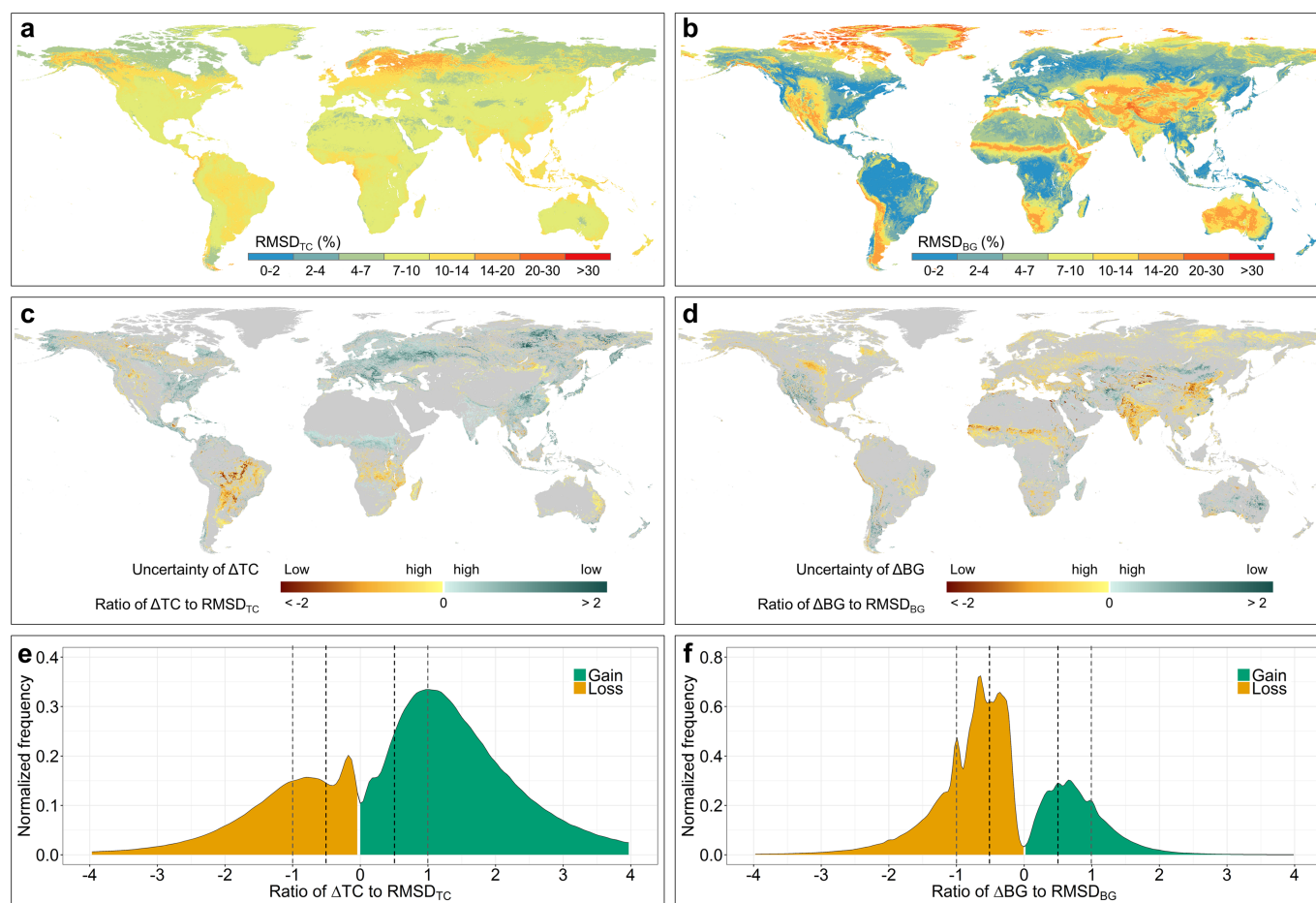
China (30.925° N, 121.175° E). **i**, Oil extraction in New Mexico, USA (32.875° N, 104.275° W). **j**, Herbaceous vegetation increase owing to glacial retreat in Chuy, Kyrgyzstan (42.575° N, 74.775° E). **k**, Bare ground cover variation along Mar Chiquita lake shore in Cordoba, Argentina (30.675° S, 63.025° W). **l**, Forest fires in Saskatchewan, Canada (55.225° N, 102.225° W). **m**, Tree cover increase in unpopulated savannahs in Western Equatoria, South Sudan^{16,17} (6.575° N, 27.725° E). **n**, Climate-change-driven woody encroachment in Quebec, Canada¹⁵ (59.475° N, 73.225° W). Examples **a–i** show various types of land use, whereas examples **j–n** do not show visible signs of human activity. Map data: Google, DigitalGlobe, CNES/Airbus, Landsat/Copernicus.

31. Gessner, U., Machwitz, M., Conrad, C. & Dech, S. Estimating the fractional cover of growth forms and bare surface in savannas. A multi-resolution approach based on regression tree ensembles. *Remote Sens. Environ.* **129**, 90–102 (2013).



Extended Data Fig. 5 | Global trends in land cover during 1982–2016. **a**, Trends in TC cover. **b**, Trends in SV cover. **c**, Trends in BG cover. The following steps were taken for each cover type using TC as the example. The TC gain layer (Fig. 1b) was overlaid on the annual TC% stack to compute annual global TC area within the gain mask (solid dark blue lines); the TC loss layer (Fig. 1b) was overlaid on the annual TC% stack

to compute annual global TC area within the loss mask (solid dark red lines). Gross gain estimates from 1986 to 2016 are marked by blue arrows and dashed lines; gross loss estimates from 1986 to 2016 are marked by red arrows and dashed lines. See Extended Data Table 1 for exact gross change estimates.



Extended Data Fig. 6 | Uncertainty of ΔTC and ΔBG . **a**, Spatial distribution of annual mean root-mean-square-deviation (RMSD) of TC between 1982 and 2016. **b**, Spatial distribution of annual mean RMSD of BG between 1982 and 2016. **c**, Spatial distribution of ΔTC uncertainty. **d**, Spatial distribution of ΔBG uncertainty. **e**, Normalized frequency distribution of ΔTC uncertainty. **f**, Normalized frequency distribution of ΔBG uncertainty. TC, BG and associated RMSD values are outputs of regression tree models. Uncertainty is represented by the ratio of long-term TC (or BG) change estimates to respective RMSD estimates. Positive

values of the ratio metric represent the uncertainties of gains and negative values represent the uncertainties of losses. A greater absolute value indicates lower uncertainty, and vice versa. Area under the frequency distribution equals 1. The frequency distributions suggest that tree cover gain exceeds tree cover loss and bare ground loss exceeds bare ground gain for any threshold level (for example, dashed lines), hence the observed trends (a net gain in tree cover and a net loss in bare ground cover over the study period) are valid.

Extended Data Table 1 | Estimates of 1982 land-cover area and 1982–2016 land-cover change at continental and global scales

Continent	Tree canopy cover							Short vegetation cover							Bare ground cover						
	Annual net change					Gross change		Annual net change					Gross change		Annual net change					Gross change	
	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)
Africa	4672	-1.9	-7.6	3.6	0.609	-267	262	11653	14.8	6.5	23.2	0.016	-268	571	13413	-12.4	-19.9	-4.7	0.020	-371	105
Asia	8457	37.5	28.0	45.3	0.000	-178	1170	21774	-22.9	-34.5	-9.6	0.008	-1261	760	13926	-15.1	-23.1	-7.4	0.002	-798	358
Europe	2719	28.3	20.4	32.8	0.000	-17	758	6320	-22.0	-27.3	-14.7	0.000	-673	50	668	-4.3	-5.8	-2.6	0.000	-92	9
North America	5815	15.6	3.5	24.2	0.020	-205	583	12921	-12.7	-4.1	-2.2	0.031	-594	286	4847	-2.5	-7.3	2.2	0.363	-186	140
South America	8767	-14.1	-20.5	-7.4	0.001	-621	190	7165	14.8	8.1	21.0	0.002	-224	655	1717	1.9	-1.2	3.8	0.307	-92	102
Oceania	680	0.1	-1.4	1.7	0.887	-40	56	4600	-4.4	-12.1	3.4	0.349	-132	50	2772	5.2	-3.9	12.5	0.280	-35	113
Global	31628	66.0	27.3	100.5	0.008	-1331	3039	64539	-26.0	-64.8	15.2	0.244	-3170	2380	37412	-34.0	-52.3	-10.0	0.023	-1582	830

Annual net change in land cover (slope) and 1982 land-cover area were estimated using Theil–Sen regression of the time series of annual land-cover area per continent or over the globe (excluding Antarctica). Lower and upper slopes represent the 90% confidence interval. Reported *P* value is for the two-sided Mann–Kendall test for trend, with *P* < 0.05 used to define statistical significance, and a sample size of *n* = 35 years. Gross change in land cover was estimated on the basis of per-pixel non-parametric trend analysis. Per-pixel loss and gain were summed to derive gross loss and gain at the aggregated scales.

Extended Data Table 2 | Estimates of 1982 land-cover area and 1982–2016 land-cover change at biome and climate zone scales

Biome / climate zone	Tree canopy cover							Short vegetation cover							Bare ground cover						
	Annual net change					Gross change		Annual net change					Gross change		Annual net change					Gross change	
	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)	Area 1982 (10 ³ km ²)	Slope (10 ³ km ² yr ⁻¹)	Lower (10 ³ km ² yr ⁻¹)	Upper (10 ³ km ² yr ⁻¹)	p	loss (10 ³ km ²)	gain (10 ³ km ²)
Tropical rainforest	10519	-1.9	-4.4	1.6	0.443	-332	315	3721	2.1	-1.0	5.1	0.307	-292	326	236	-0.4	-0.7	-0.1	0.025	-19	15
Tropical moist deciduous forest	3569	-2.5	-8.5	2.2	0.460	-373	285	6912	5.8	0.3	10.9	0.078	-236	386	492	-2.2	-3.0	-1.3	0.001	-71	29
Tropical dry forest	1236	-2.8	-5.1	-1.1	0.018	-184	99	5386	7.2	4.0	10.2	0.001	-70	246	821	-3.8	-5.9	-1.9	0.010	-121	32
Tropical mountain system	1333	3.5	2.4	4.5	0.000	-23	118	2092	-1.4	-3.1	0.2	0.118	-106	65	1092	-1.7	-2.7	-0.9	0.002	-61	17
Tropical shrubland	149	0.3	-0.2	0.7	0.349	-15	20	4010	12.3	6.8	18.5	0.001	-41	371	4137	-12.0	-19.0	-6.1	0.003	-379	43
Tropical desert	19	0.0	0.0	0.0	0.532	0	1	692	1.6	0.2	3.4	0.061	-31	87	10846	-1.5	-3.4	-0.1	0.057	-88	31
Tropical climate zone	16837	-4.1	-14.4	3.6	0.320	-927	837	22691	30.0	14.7	43.0	0.002	-775	1480	17617	-25.5	-34.7	-12.0	0.002	-740	167
Subtropical humid forest	1566	8.2	4.4	12.0	0.002	-48	268	2866	-7.3	-10.5	-3.7	0.003	-236	46	196	-0.7	-1.4	-0.3	0.012	-38	22
Subtropical mountain system	516	3.1	2.3	3.8	0.000	-20	116	2571	-2.8	-4.3	-1.2	0.008	-153	68	1756	0.0	-1.7	1.8	0.932	-79	79
Subtropical dry forest	198	1.6	0.9	2.3	0.001	-8	49	1107	0.2	-0.5	0.8	0.755	-37	33	266	-1.2	-1.9	-0.6	0.002	-45	10
Subtropical steppe	179	-0.8	-2.0	0.2	0.191	-27	12	2594	-1.4	-4.2	1.8	0.460	-84	64	2106	3.2	-0.9	6.7	0.201	-61	106
Subtropical desert	29	-0.2	-0.4	0.0	0.118	-3	2	2606	-4.4	-10.3	1.3	0.233	-128	45	4001	4.5	-1.3	10.6	0.233	-46	133
Subtropical climate zone	2453	12.1	6.5	16.8	0.004	-105	448	11741	-14.0	-23.1	-5.6	0.013	-639	257	8323	5.6	-6.8	17.0	0.443	-269	350
Temperate continental forest	2172	21.4	15.1	26.0	0.000	-11	591	4451	-17.9	-22.5	-11.3	0.000	-528	28	277	-2.7	-3.4	-2.2	0.000	-61	7
Temperate mountain system	1552	5.9	3.6	7.4	0.001	-53	198	3459	-2.0	-4.0	0.5	0.211	-213	172	2175	-2.9	-5.1	-1.1	0.023	-161	62
Temperate oceanic forest	551	3.8	2.1	5.3	0.001	-6	101	1162	-3.3	-4.9	-1.8	0.003	-92	8	61	-0.4	-0.5	-0.2	0.000	-8	2
Temperate steppe	320	2.2	0.1	3.4	0.069	-18	56	4191	-2.9	-6.7	0.0	0.105	-130	72	1338	2.3	-1.5	5.8	0.363	-86	108
Temperate desert	61	-0.1	-0.2	0.1	0.514	-5	4	1661	-0.3	-2.9	3.5	0.955	-101	135	3642	0.3	-3.6	3.1	0.887	-135	103
Temperate climate zone	4681	33.5	21.0	41.9	0.000	-92	951	14814	-24.3	-37.3	-12.2	0.006	-1064	414	7491	-4.3	-14.0	1.8	0.268	-451	282
Boreal coniferous forest	3938	13.6	6.7	18.7	0.003	-75	415	4239	-12.6	-17.1	-6.3	0.002	-369	71	205	-1.4	-1.8	-0.9	0.001	-23	2
Boreal mountain system	2035	6.6	3.9	9.9	0.005	-61	225	3909	-6.2	-8.8	-3.3	0.003	-193	64	341	-1.2	-1.7	-0.6	0.002	-33	11
Boreal tundra woodland	971	1.3	-1.5	3.7	0.363	-58	82	2723	-0.9	-2.9	1.2	0.478	-63	52	228	-0.7	-1.2	-0.2	0.044	-16	5
Boreal climate zone	6796	21.0	7.9	31.0	0.009	-194	723	10857	-20.3	-27.5	-11.7	0.002	-625	187	772	-3.4	-4.5	-2.2	0.001	-71	19
Polar	236	2.2	0.9	3.1	0.009	-7	55	4109	0.4	-1.3	2.1	0.712	-43	30	3080	-2.6	-3.6	-1.1	0.010	-41	8

Consistent with Extended Data Table 1, annual net change in land cover (slope) and 1982 land-cover area were estimated using Theil–Sen regression of the time series of annual land-cover area per biome or climate zone. Lower and upper slopes represent the 90% confidence interval. Reported *P* value is for the two-sided Mann–Kendall test for trend with $P < 0.05$ used to define statistical significance and a sample size of $n = 35$ years. Gross change in land cover was estimated on the basis of per-pixel non-parametric trend analysis. Per-pixel loss and gain were summed to derive gross loss and gain at the aggregated scales. See Extended Data Fig. 2q for the geographical distribution of biomes.

A multi-cohort study of the immune factors associated with *M. tuberculosis* infection outcomes

Roshni Roy Chowdhury^{1,2}, Francesco Vallania^{3,4}, Qiantian Yang⁵, Cesar Joel Lopez Angel^{1,2}, Fatoumatta Darboe^{6,7}, Adam Penn-Nicholson^{6,7}, Virginie Rozot^{6,7}, Elisa Nemes^{6,7}, Stephanus T. Malherbe^{8,9}, Katharina Ronacher^{9,10,11}, Gerhard Walzl^{9,10}, Willem Hanekom^{6,12}, Mark M. Davis^{1,2,3,13}, Jill Winter⁸, Xinchun Chen¹⁴, Thomas J. Scriba^{6,7}, Purvesh Khatri^{2,3,4*} & Yueh-hsiu Chien^{1,2*}

Most infections with *Mycobacterium tuberculosis* (*Mtb*) manifest as a clinically asymptomatic, contained state, known as latent tuberculosis infection, that affects approximately one-quarter of the global population¹. Although fewer than one in ten individuals eventually progress to active disease², tuberculosis is a leading cause of death from infectious disease worldwide³. Despite intense efforts, immune factors that influence the infection outcomes remain poorly defined. Here we used integrated analyses of multiple cohorts to identify stage-specific host responses to *Mtb* infection. First, using high-dimensional mass cytometry analyses and functional assays of a cohort of South African adolescents, we show that latent tuberculosis is associated with enhanced cytotoxic responses, which are mostly mediated by CD16 (also known as FcγRIIIa) and natural killer cells, and continuous inflammation coupled with immune deviations in both T and B cell compartments. Next, using cell-type deconvolution of transcriptomic data from several cohorts of different ages, genetic backgrounds, geographical locations and infection stages, we show that although deviations in peripheral B and T cell compartments generally start at latency, they are heterogeneous across cohorts. However, an increase in the abundance of circulating natural killer cells in tuberculosis latency, with a corresponding decrease during active disease and a return to baseline levels upon clinical cure are features that are common to all cohorts. Furthermore, by analysing three longitudinal cohorts, we find that changes in peripheral levels of natural killer cells can inform disease progression and treatment responses, and inversely correlate with the inflammatory state of the lungs of patients with active tuberculosis. Together, our findings offer crucial insights into the underlying pathophysiology of tuberculosis latency, and identify factors that may influence infection outcomes.

Although most *Mtb* infections do not lead to the manifestation of clinical disease, few studies have focused on delineating the immune factors that are associated with the asymptomatic states that comprise latent tuberculosis infection (LTBI). To broadly characterize this immune state, we used high-dimensional cytometry by time-of-flight (CyTOF), a proteomics technology that assesses the abundance of cell subsets, protein expression and activation of signalling pathways at the single-cell resolution⁴ (Fig. 1a). We analysed peripheral blood mononuclear cells (PBMCs) from uninfected and latently infected adolescents (aged 13–18 years) from South Africa (Supplementary Table 1). This cohort is from a highly endemic area but has a lower rate of active tuberculosis (TB) than is seen in young children and adults⁵, indicating a well-controlled *Mtb* infection.

An initial analysis (Supplementary Table 2) of 14 uninfected controls and 14 individuals with LTBI identified four cell subsets (defined by cell-surface protein expression) with a significantly higher percentage (of total live cells) in individuals with LTBI than uninfected controls (false discovery rate (FDR) of <1%). These four subsets comprised total CD16-expressing cells, natural killer (NK) cells and two closely related populations of CD27⁺CD8⁺αβ T cells that differed in their CD38 expression. By contrast, two other cell subsets, total B cells and naive B cells, were significantly less abundant in individuals with LTBI (Fig. 2a and Extended Data Fig. 1a–c). Similar differences in NK cell and B cell percentages between uninfected controls and individuals with LTBI were also observed in an additional 20 individuals analysed by CyTOF, and another 32 individuals analysed by flow cytometry (Extended Data Fig. 1d). Because latently infected individuals show no significant change in peripheral monocyte or lymphocyte counts compared to uninfected controls^{6,7}, changes in the percentage of a given cell type most likely reflect corresponding alterations in its abundance.

Significant differences in the percentage of immune effector cell subsets between samples from uninfected controls and individuals with LTBI were also identified. Granzyme B (GZMB) and perforin (PRF) expressing cells were significantly higher in individuals with LTBI. These mostly consisted of NK cells and GZMB⁺PRF⁺IFNγ⁺TNF⁺ polyfunctional cells, which largely comprised CD27⁺CD8⁺αβ T cells, but also included NK cells and γδ T cells (Fig. 2b and Extended Data Fig. 2a–c). These cells also expressed significantly higher levels of GZMB (Extended Data Fig. 2d, e), indicating an enhanced cytotoxic potential on a per-cell basis. Indeed, NK cells from individuals with LTBI showed significantly higher target cell lysis than those from uninfected controls ($n = 10$ per group, $P = 0.003$; Fig. 2c). Additionally, there were higher percentages of CD16⁺GZMB^{high} cells within the compartments of the NK cells, CD8⁺αβ T cells and γδ T cells in PBMCs from individuals with LTBI (Fig. 2d and Extended Data Fig. 2f). PBMCs from individuals with LTBI also mounted significantly higher antibody-dependent cell-mediated cytotoxicity (ADCC) responses than those from uninfected controls ($n = 12$ per group, $P = 0.006$; Fig. 2e and Extended Data Fig. 2g). ADCC allows antibodies, in addition to T cells, to contribute to the antigen-specific cytotoxic response. In this context, it was reported that antibodies from individuals with LTBI, compared to those from patients with active TB, have unique Fc functional profiles that promote selective binding to CD16 and effectively drive intracellular *Mtb* killing⁸.

The ability of cells to respond to immune challenges is an important factor in maintaining host immune competence. To determine the signalling

¹Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. ²Program in Immunology, Stanford University School of Medicine, Stanford, CA, USA. ³Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, USA. ⁴Division of Biomedical Informatics, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. ⁵Shenzhen Key Laboratory of Infection and Immunity, Shenzhen Third People's Hospital, Shenzhen, China. ⁶South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. ⁷Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa. ⁸Catalysis Foundation for Health, Emeryville, CA, USA. ⁹Department of Science and Technology, National Research Foundation Centre of Excellence for Biomedical Tuberculosis Research, Stellenbosch University, Stellenbosch, South Africa. ¹⁰South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Stellenbosch, South Africa. ¹¹Mater Research Institute, The University of Queensland, Brisbane, Queensland, Australia. ¹²Department of Pediatrics and Child Health, University of Cape Town, Cape Town, South Africa. ¹³Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA. ¹⁴Department of Pathogen Biology, Shenzhen University School of Medicine, Shenzhen, China. *e-mail: pkhatri@stanford.edu; chien@stanford.edu

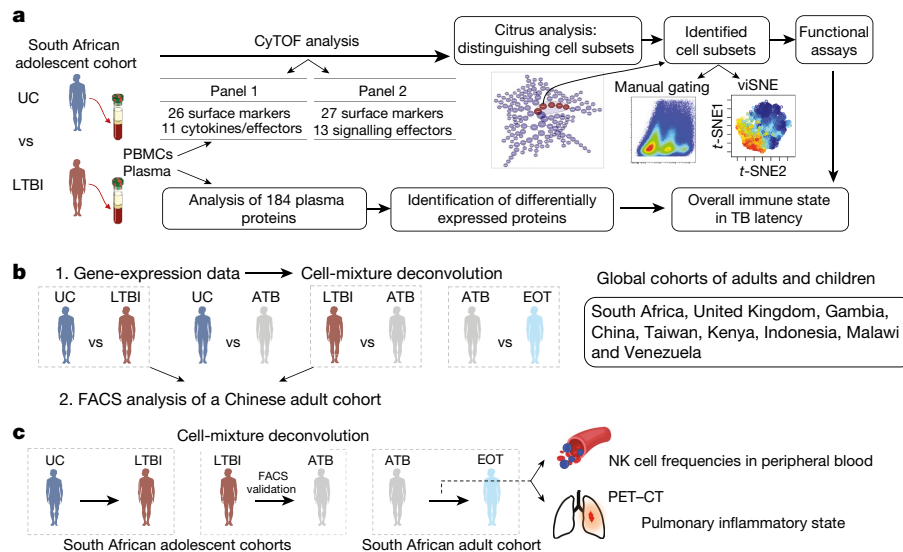


Fig. 1 | Schematic representation of the experimental design.

a, Identification of immune features distinguishing uninfected and latently infected individuals from a cohort of South African adolescents. *t*-SNE, *t*-distributed stochastic neighbour embedding. viSNE, visualization using *t*-SNE. **b**, Analysis of changes in immune cell subset abundance at different stages of infection and end-of-treatment using cell-type deconvolution of transcriptomic data from multiple cohorts, and fluorescence-activated cell

sorting (FACS) analysis of PBMCs from an adult Chinese cohort.

c, Evaluation of changes in NK cell frequencies in longitudinal cohorts, for individuals who (1) acquired *Mtb* infection (QuantiFERON converters); (2) progressed from LTBI to active TB, and (3) patients with active TB who proceeded to treatment completion; and their correlations with pulmonary pathology as measured by PET-CT imaging. ATB, active tuberculosis; EOT, end-of-treatment; UC, uninfected controls.

capacity of immune cells, we used CyTOF (Supplementary Table 2) to investigate their ability to transiently phosphorylate signalling effectors in response to PMA and ionomycin, IFN γ , TNF or combined anti-CD3 and anti-CD28 stimulation. We found that in individuals with LTBI, all T cell subsets exhibited diminished responsiveness through the S6 signalling

pathway, irrespective of the stimulation condition, with the exception of $\gamma\delta$ T cells after stimulation with both anti-CD3 and anti-CD28, as compared to cells from uninfected controls (Fig. 2f). S6, a ribosomal component, is phosphorylated after mTOR activation. This signalling pathway is critical for ribosome biogenesis, cell growth and proliferation⁹.

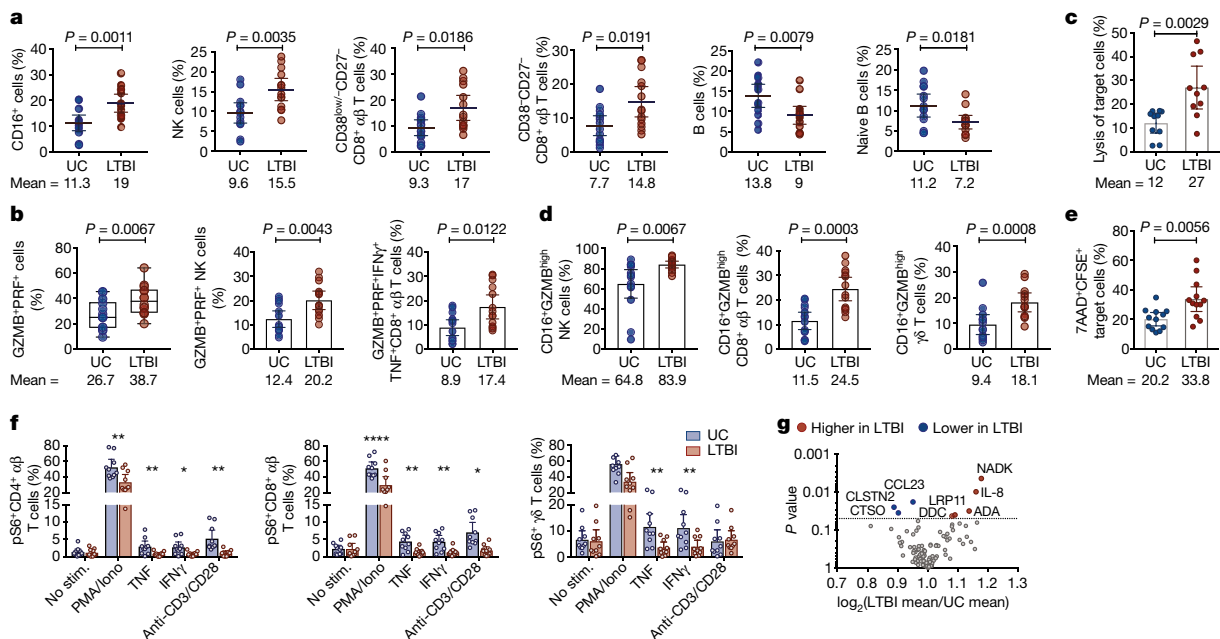


Fig. 2 | Immune state of TB latency identified in a cohort of South African adolescents. **a**, Frequencies of cell subsets were defined by surface marker (**a**) and effector molecule (**b**) expression that are present in significantly ($FDR < 1\%$ by SAM analysis) different abundances between uninfected controls and individuals with LTBI ($n = 14$ per group) as determined by Citrus analysis of CyTOF results (Extended Data Figs. 1, 2). **c**, Cytolytic responses of NK cells isolated from PBMCs of uninfected controls and individuals with LTBI ($n = 10$ per group), quantified by calcein-release from calcein-labelled target (K562) cells upon lysis. **d**, Percentages of CD16⁺GZMB^{high} cells within each lymphocyte subset in uninfected controls and individuals with LTBI ($n = 14$ per group)

(Extended Data Fig. 2f). **e**, ADCC response of total PBMCs from uninfected controls and individuals with LTBI ($n = 12$ per group) as determined by antibody-mediated killing of CFSE-labelled target (P815) cells (Extended Data Fig. 2g). **f**, Frequencies of phosphorylated ribosomal protein S6 (pS6)⁺ cells within T cell subsets under different stimulation conditions in uninfected controls and individuals with LTBI ($n = 10$ per group). **g**, Volcano plot of plasma protein abundance in uninfected controls and individuals with LTBI ($n = 27$ per group) (Supplementary Table 3). Throughout, P values were derived using a Mann–Whitney U -test, unless otherwise stated. Mean and error bars representing the 95% confidence intervals are shown for each comparison. See Supplementary Table 1.

Although alterations in the T cell compartment in individuals with LTBI did not lead to alteration in the levels of T cells in the periphery (Extended Data Fig. 1e), changes in the B cell compartment resulted from a decrease in the abundance of total B cells, which was largely driven by a reduction in circulating naive B cells. Reductions in the levels of peripheral B cells could be due to preferential sequestration of these cells at the site of infection (lungs and associated lymph nodes) and/or altered output of B cells from the bone marrow; inflammation can lead to enhanced myelopoiesis with diminished B cell output¹⁰. Analysis of 184 plasma proteins (Supplementary Table 3) showed significantly higher levels of inflammation-associated molecules, such as CXCL8 (also known as IL-8; $P = 0.01$), adenosine deaminase (ADA; $P = 0.035$) and NAD kinase (NADK; $P = 0.004$) in samples from individuals with LTBI relative to uninfected controls ($n = 27$ per group). By contrast, the plasma levels of CCL23, which has been shown to inhibit myelopoiesis¹¹, was significantly lower in individuals with LTBI ($P = 0.02$; Fig. 2g and Extended Data Fig. 3), suggesting altered myelopoiesis in LTBI. Taken together, our results identified multiple immune components that operate together in LTBI. Specifically, in the presence of ongoing inflammation coupled with deviations in B and T cell compartments, enhanced cytotoxic responses that are mostly mediated by CD16 and NK cells appeared to be key factors associated with maintaining latency, which we propose represents successful immune control of *Mtb* infection.

Because alterations in peripheral immune cell distributions have not previously been commonly associated with TB latency, we tested whether such changes were observed in other LTBI cohorts, including those of children and adults (Fig. 1b). Although their PBMCs were not available for analysis, transcriptional profiles of whole-blood or PBMC samples from these cohorts were publicly available. We applied a computational approach to infer leukocyte representations from gene-expression profiles using support vector regression¹² with the leukocyte expression signature matrix ‘immunoStates’¹³. Analysis of gene-expression datasets from 189 uninfected controls and 145 subjects with LTBI from six clinical cohorts (Supplementary Table 4), including children and adults from four continents, showed that NK cells were significantly more abundant (FDR = 0.018%) in cohorts of individuals with LTBI (Fig. 3a). Changes in B cell percentages were heterogeneous across the cohorts. However, analysis from all cohorts combined indicated a significant reduction in the percentages of total B cells (FDR = 1%) and naive B cells (FDR = 0.38%) in samples from individuals with LTBI (Fig. 3a and Extended Data Fig. 4a). No significant differences in total T cell abundance (FDR = 74%) (Fig. 3a) or those of the CD4⁺ and CD8⁺ subsets (FDR = 44% for both subsets; Extended Data Fig. 5a) were observed between the uninfected and LTBI cohorts. Thus, the analyses from multiple LTBI cohorts were consistent with our findings obtained from the cohort of South African adolescents.

To evaluate whether and how these immune cell frequencies change during active disease, we analysed transcriptome profiles from PBMC or whole-blood samples from 409 individuals with LTBI and 543 patients with active TB from nine clinical cohorts (Supplementary Table 4), profiling children and adults, with and without HIV comorbidity, across three continents. Deconvolution estimations showed that NK cells were present at significantly (FDR = 0%, $P < 2.2 \times 10^{-16}$) lower frequencies in samples from patients with active TB, with notable consistency across all cohorts (Fig. 3b). Flow cytometry analysis of whole-blood samples from adults collected from 24 uninfected controls, 17 individuals with LTBI and 23 patients with active TB (Supplementary Table 5) from a non-endemic region (Shenzhen) in China also confirmed the stage-specific changes in NK cell frequencies (Fig. 3d).

Our deconvolution estimations also showed significantly reduced levels of B cells in patients with active TB compared to uninfected individuals (Extended Data Fig. 6). The analysis of B cells in an adult Italian cohort is one such example⁷. In general, the decrease in the frequencies of B cells and naive B cells was observed in either latency or active disease relative to uninfected individuals (Fig. 3a, b and Extended

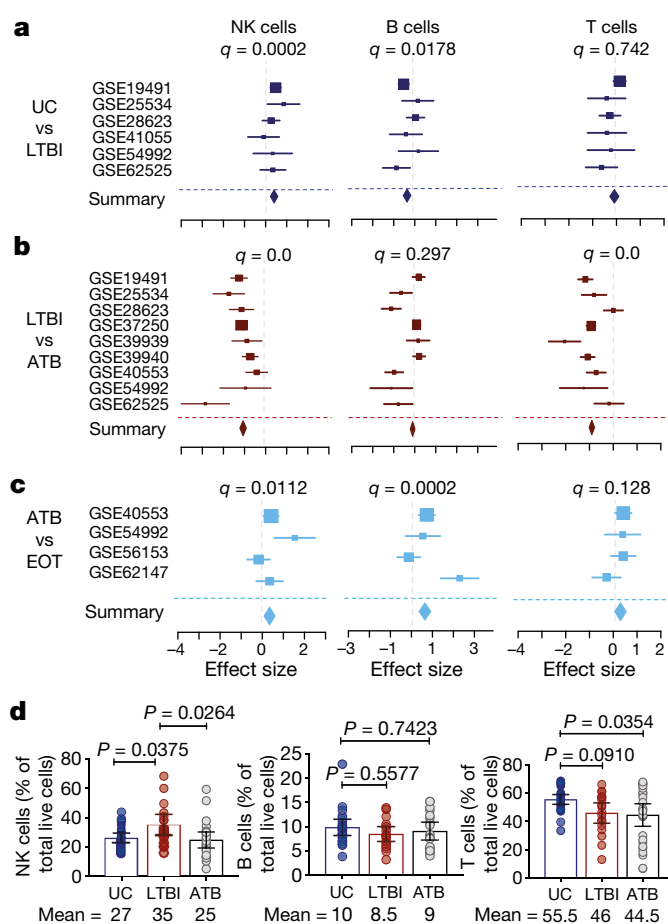


Fig. 3 | Peripheral lymphocyte distributions at different infection stages from global cohorts. Forest plots comparing changes in the levels of NK cells, B cells and T cells were calculated using cell-mixture deconvolution. **a**, Comparison between uninfected controls ($n = 189$) and individuals with LTBI ($n = 145$). **b**, Comparison between individuals with LTBI ($n = 409$) and patients with active TB ($n = 543$). **c**, Comparison between patients with active TB before ($n = 76$) and after six months of treatment (end-of-treatment) ($n = 97$). Cohort GSE identifiers are listed on the left. Boxes represent the standardized mean difference in estimated cellular proportions in a cohort between two comparison groups (effect size). The size of the box is proportional to the sample size of a given cohort. Whiskers represent the 95% confidence interval of the corresponding standardized mean difference in cellular proportions. Diamonds represent the overall difference in cellular proportions between two groups by integrating the standardized mean differences across all individual cohorts-summary effect sizes (Summary). The width of the diamond corresponds to its 95% confidence interval. The q values (FDR) for the summary effect sizes are shown above each plot. **d**, Percentages of peripheral NK cells, B cells and T cells in a Chinese cohort of uninfected controls ($n = 24$), individuals with LTBI ($n = 17$) and patients with active TB ($n = 23$) assessed by flow cytometry. P values were derived using a one-way ANOVA with Tukey's multiple comparisons test. Mean and error bars representing the 95% confidence intervals are shown for each comparison. See Supplementary Tables 4, 5.

Data Fig. 4a, b), with the exception of an Indonesian cohort, in which no change in B cell frequencies was observed. Notably, similar to the Indonesian cohort, no change in B cell frequencies was observed in the adult Chinese cohort, assessed by flow cytometry (Fig. 3d). These observations underscore the heterogeneity of immune states in *Mtb* infection. Although our South African adolescent cohort showed no significant changes in T cell abundance between uninfected controls and individuals with LTBI, both CD4⁺ and CD8⁺ $\alpha\beta$ T cell populations showed significantly diminished responsiveness through the S6 pathway, which may foretell the eventual drop in peripheral T cell levels observed in patients with active TB (Fig. 3b, Extended Data Fig. 5b).

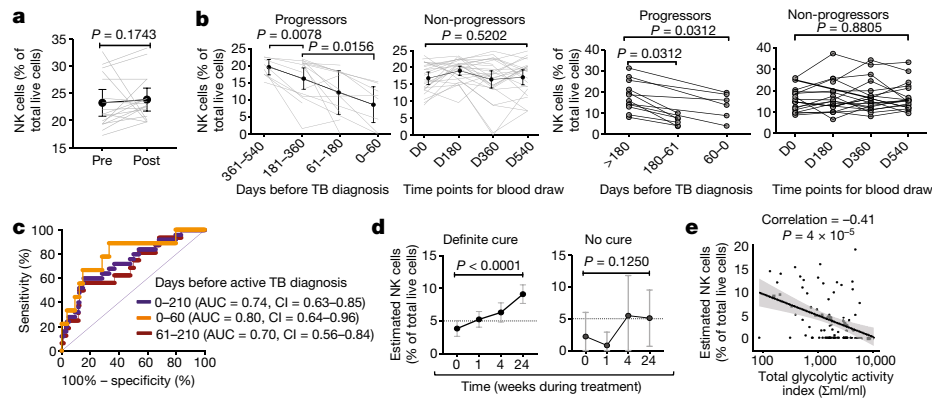


Fig. 4 | Correlations between peripheral NK cell percentages and disease progression, treatment response and inflammation in the lung. **a**, Changes in peripheral NK cell percentages in South African adolescents after acquisition of *Mtb* infection (Extended Data Fig. 9, $n = 17$) were determined by cell-mixture deconvolution. Pre-infection (Pre) gene-expression data (180–360 days) were compared to data obtained at the time of infection diagnosis or the nearest time point after diagnosis (0–360 days) (Post). **b**, Changes in peripheral NK cell percentages during progression from LTBI to active disease at different time points before TB diagnosis and non-progressors over a span of two years (Supplementary Table 6) were determined by cell-mixture deconvolution (17 progressors and 26 non-progressors) (left) and flow cytometry (12 progressors and 20 non-progressors) (right). All P values were derived using a Wilcoxon rank-sum test. Mean and error bars representing the 95% confidence intervals are shown. **c**, Receiver operating characteristic curves of the

potential of estimated NK cell frequencies as a predictor of TB disease progression. **d**, Estimated NK cell percentages in patients with active TB from the Catalysis-TB cohort at baseline (pre-treatment) and at various time points during treatment. Definite cure indicates sputum culture negative by month 6 after treatment initiation ($n = 76$); no cure indicates sputum culture positive after six months of treatment initiation ($n = 7$). P values were derived using a Wilcoxon rank-sum test. Mean and error bars representing the 95% confidence intervals are shown. **e**, Correlation plot showing the relationship between estimated peripheral NK cell frequencies in patients with active TB at baseline (pre-treatment) and total glycolytic activity index (TGAI) measured by PET-CT imaging of the lungs at baseline. The line represents the best fit and the shaded area the 95% confidence interval. NK cell frequencies were determined by deconvolution.

Flow cytometry assessment of the Chinese cohort confirmed the significant decrease in peripheral T cell levels in patients with active TB compared to uninfected controls (Fig. 3d). Whether reduced T cell signalling capacity is a common feature of LTBI requires further analysis. Nonetheless, our results suggest that immune deviations in both T and B cell compartments start in latency and progress further in active disease.

To test whether the disease-induced reductions in peripheral lymphocyte populations recover after successful treatment, we deconvoluted PBMC or whole-blood transcriptome profiles of 76 samples from patients with active TB and 97 samples from patients at the end of treatment from four independent cohorts of HIV-negative adults (Supplementary Table 4) from three continents. NK cell frequencies, together with all major immune cell populations, except $CD4^+ \alpha\beta$ T cells, were significantly higher in successfully treated individuals relative to patients with active TB, reaching reference levels observed in healthy, *Mtb*-uninfected individuals (Fig. 3c and Extended Data Fig. 4c, 5c). A longer recovery time for $CD4^+ \alpha\beta$ T cells might be indicative of a post-treatment inflammatory state due to ongoing subclinical disease as seen by positron emission tomography and computed tomography (PET-CT) analysis¹⁴. The trajectories of estimated immune cell frequencies through the different stages of infection and after treatment are summarized in Extended Data Fig. 7.

Because of the apparent correlation between changes in NK cell frequencies and the stages of *Mtb* infection, we tested whether the levels of peripheral NK cells could inform TB disease progression and response to treatment (Fig. 1c) by analysing three independent longitudinal follow-up studies of South African cohorts, which included individuals who (1) acquired latent *Mtb* infection (defined as converting from QuantiFERON (QFT)-negative to QFT-positive), (2) progressed from latent infection to active disease (South African adolescent progressor cohort)^{15,16} and (3) proceeded from active disease to treatment completion (Catalysis-TB cohort)¹⁷. Deconvolution estimations showed that relative to pre-infection, levels of NK cells did not change significantly after acquisition of *Mtb* infection ($n = 17$; Fig. 4a) and decreased during the progression from latent infection to active disease ($n = 17$; Fig. 4b (left)). Consistent with this latter computational finding, flow cytometry analysis of 32 individuals with

LTBI (12 progressors and 20 non-progressors) (Supplementary Table 6) showed that NK cell frequencies decreased in each of the progressors, 0–180 days before TB diagnosis, whereas the non-progressors showed no significant change in peripheral NK cell frequencies during the two-year study period (Fig. 4b (right)). Statistical analysis of the predictive power of NK cell levels for progression to active disease using receiver operating characteristic curves showed an area under the curve of 0.74 (95% confidence interval 0.63–0.85) in the seven months preceding TB diagnosis (Fig. 4c). In addition, deconvolution analysis of gene-expression data from the Catalysis Foundation for Health study of patients with active TB under treatment showed that in individuals who responded to treatment (classified as ‘definite cure’, $n = 76$), NK cell levels were significantly higher at the end of treatment (week 24) compared to baseline (pre-treatment, $P < 0.0001$; Fig. 4d). By contrast, treatment non-responders (classified as ‘no cure’, $n = 7$), showed no significant change in their NK cell percentages between baseline and end of treatment ($P = 0.1250$; Fig. 4d). Furthermore, we found that the inflammatory burden of the lung indicated by the total glycolytic activity index, as measured by PET-CT¹⁴, inversely correlated with peripheral NK cell frequencies at diagnosis (pre-treatment; Fig. 4e) and at week 4 after treatment initiation (Extended Data Fig. 8). Therefore, changes in peripheral NK cell levels reflect changes in the activity level and burden of *Mtb* in the lung. These observations support the view that circulating NK cells reflect key features of the host immune state, can serve as surrogates of the immune response at the nidus of infection and that longitudinal measurements of peripheral NK cells can inform disease progression and treatment efficacy.

NK cells have been shown to kill *Mtb*-infected cells directly or through ADCC¹⁸. Moreover, NK cells become activated and expand in the lung during the early response in a mouse model of aerosol exposure with *Mtb*, but depleting NK cells from these immunocompetent mice does not alter the course of infection¹⁹. Nevertheless, in mice with T cell deficiencies, NK cells were found to confer protection against *Mtb* infection²⁰. Although mice infected with *Mtb* do not establish latency, these observations are consistent with the proposition that in the immune state that we observed here in TB latency, NK cells could contribute to protective immunity. Along this line, TB is the most common fatal opportunistic infection in HIV/AIDS²¹, and NK cells are

noted for controlling HIV infections²². Progressive impairment of NK cell functions and depletion of NK cells, especially the CD16⁺ subsets, have been noted in HIV infection²³. Additionally, anti-TNF therapy, which is associated with increased incidence of TB disease in auto-immune patients with LTBI²⁴, was shown to impair NK cell function²⁵ and reduce the expression of PRF and granzyme in lymphocytes²⁶.

Taken together, our analyses offer a better understanding of the immune state of latent *Mtb* infection and factors that mediate and/or predict transitions from latent infection to active disease. These findings may be useful for generating hypotheses that could lead to new intervention strategies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0439-x>.

Received: 18 November 2017; Accepted: 9 July 2018;

Published online 22 August 2018.

- Houben, R. M. & Dodd, P. J. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Med.* **13**, e1002152 (2016).
- Shea, K. M., Kammerer, J. S., Winston, C. A., Navin, T. R. & Horsburgh, C. R. Jr. Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup. *Am. J. Epidemiol.* **179**, 216–225 (2014).
- WHO. *Global Tuberculosis Report*. http://www.who.int/tb/publications/global_report/en/ (2017).
- Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165**, 780–791 (2016).
- Mahomed, H. et al. Predictive factors for latent tuberculosis infection among adolescents in a high-burden area in South Africa. *Int. J. Tuberc. Lung Dis.* **15**, 331–336 (2011).
- Takenami, I. et al. Blood cells and interferon-gamma levels correlation in latent tuberculosis infection. *ISRN Pulmonol.* **2013**, 256148 (2013).
- Joosten, S. A. et al. Patients with tuberculosis have a dysfunctional circulating B-cell compartment, which normalizes following successful treatment. *PLoS Pathog.* **12**, e1005687 (2016).
- Lu, L. L., et al. A functional role for antibodies in tuberculosis. *Cell* **167**, 433–443 (2016).
- Ruvinsky, I. & Meyuhas, O. Ribosomal protein S6 phosphorylation: from protein synthesis to cell size. *Trends Biochem. Sci.* **31**, 342–348 (2006).
- Ueda, Y., Kondo, M. & Kelsoe, G. Inflammation and the reciprocal production of granulocytes and lymphocytes in bone marrow. *J. Exp. Med.* **201**, 1771–1780 (2005).
- Shih, C. H., van Eeden, S. F., Goto, Y. & Hogg, J. C. CCL23/myeloid progenitor inhibitory factor-1 inhibits production and release of polymorphonuclear leukocytes and monocytes from the bone marrow. *Exp. Hematol.* **33**, 1101–1108 (2005).
- Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).
- Vallania, F. et al. Leveraging heterogeneity across multiple data sets increases accuracy of cell-mixture deconvolution and reduces biological and technical biases. Preprint at <https://biorxiv.org/content/early/2017/10/20/206466> (2017).
- Malherbe, S. T. et al. Persisting positron emission tomography lesion activity and *Mycobacterium tuberculosis* mRNA after tuberculosis cure. *Nat. Med.* **22**, 1094–1100 (2016).
- Zak, D. E. et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet* **387**, 2312–2322 (2016).
- Scriba, T. J. et al. Sequential inflammatory processes define human progression from *M. tuberculosis* infection to tuberculosis disease. *PLoS Pathog.* **13**, e1006687 (2017).
- Thompson, E. G. et al. Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* **107**, 48–58 (2017).
- Esin, S. & Batoni, G. Natural killer cells: a coherent model for their functional role in *Mycobacterium tuberculosis* infection. *J. Innate Immun.* **7**, 11–24 (2015).
- Junqueira-Kipnis, A. P. et al. NK cells respond to pulmonary infection with *Mycobacterium tuberculosis*, but play a minimal role in protection. *J. Immunol.* **171**, 6039–6045 (2003).
- Feng, C. G. et al. NK cell-derived IFN- γ differentially regulates innate resistance and neutrophil response in T cell-deficient hosts infected with *Mycobacterium tuberculosis*. *J. Immunol.* **177**, 7086–7093 (2006).
- Kwan, C. K. & Ernst, J. D. HIV and tuberculosis: a deadly human syndemic. *Clin. Microbiol. Rev.* **24**, 351–376 (2011).
- Scully, E. & Alter, G. NK cells in HIV disease. *Curr. HIV/AIDS Rep.* **13**, 85–94 (2016).
- Mansour, I., Doinel, C. & Rouger, P. CD16⁺ NK cells decrease in all stages of HIV infection through a selective depletion of the CD16⁺CD8⁺CD3⁺ subset. *AIDS Res. Hum. Retroviruses* **6**, 1451–1457 (1990).
- Xie, X., Li, F., Chen, J. W. & Wang, J. Risk of tuberculosis infection in anti-TNF- α biological therapy: from bench to bedside. *J. Microbiol. Immunol. Infect.* **47**, 268–274 (2014).
- Nocturne, G. et al. Impact of anti-TNF therapy on NK cells function and on immunosurveillance against B-cell lymphomas. *J. Autoimmun.* **30**, 56–64 (2017).
- Bruns, H. et al. Anti-TNF immunotherapy reduces CD8⁺ T cell-mediated antimicrobial activity against *Mycobacterium tuberculosis* in humans. *J. Clin. Invest.* **119**, 1167–1177 (2009).

Acknowledgements We thank R.-P. Sekaly, A. Filali-Mouhim and K. Ghneim for transcriptional analysis of the *Mtb* acquisition subcohort of the Adolescent Cohort Study; E. Long, C. Blish for advice and the P815 mouse cell line and K562 human cell line; A. Kasmar for critically reading the manuscript. This work was supported by the Bill and Melinda Gates Foundation (T.J.S., P.K., Y.-h.C.), the National Institutes of Health AI127128 (Y.-h.C.), AI109662, AI057229 and AI125197 (P.K.), K12 5K12HL120001 (F.V.), 5T32AI07290-31 (R.R.C.), VirBio (P.K.), the National Science Foundation of China (81525016, 81772145) and the Science and Technology Project of Shenzhen (JSGG20160427104724699) (X.C.).

Reviewer information Nature thanks T. H. M. Ottenhoff and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.R.C. and Y.-h.C. conceptualized the study; F.V. and P.K. established immunoStates and performed cell-type deconvolution of transcriptome data. R.R.C. performed experiments and analysis of South African cohorts and overall data analysis. F.D., A.P.-N., V.R., S.T.M. and T.J.S. provided samples and analysed South African cohorts of individuals who progressed from LTBI to active disease. Q.Y. and X.C. performed FACS analysis of the Chinese cohort. T.J.S. and E.N. designed and oversaw the *Mtb* acquisition subcohort of the Adolescent Cohort Study. S.T.M., K.R., G.W. and J.W. provided lung pathology results from the Catalysis cohort. T.J.S. and W.H. established the South African Tuberculosis cohorts. M.M.D., C.J.L.A. and T.J.S. contributed to project design and interpretation. R.R.C. and Y.-h.C. wrote the manuscript with input from T.J.S., P.K., J.W., F.V., M.M.D. and A. P.-N. P.K. designed and oversaw all computational analyses. Y.-h.C. supervised the study.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0439-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0439-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.K. or Y.-h.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study design and participants. A South African adolescent cohort (Adolescent Cohort Study (ACS))⁵, aged 13–18 years, who were either uninfected or latently infected with *Mtb* (LTBI) (Supplementary Table 1), were analysed to characterize the immune state of TB latency. All adolescents whose parents or legal guardians provided written, informed consent and who provided written, informed assent themselves were enrolled. The study protocols were approved by the relevant human research ethics committees. Individuals were classified as latently infected if diagnosed positive by a QuantiFERON TB Gold In-tube assay (Qiagen; >0.35 IU ml⁻¹). All participants were healthy without signs or symptoms of active disease. Only adolescents who remained disease-free for two years from the time of enrolment were included in the analysis.

Cohorts associated with publicly available datasets^{27–37} that were used for the analysis of immune cell distributions at different stages of *Mtb* infection, and end-of-treatment, by cell-mixture deconvolution, are described in Supplementary Table 4.

An adult Chinese cohort (Supplementary Table 5) from Shenzhen, a non-endemic region in China was assessed for immune cell distributions at different stages of *Mtb* infection using flow cytometry analysis of PBMC samples. Individuals were defined as latently infected if diagnosed positive by the interferon gamma release assay (IGRA⁺) but showed no symptom or chest X-ray signs suggestive of active disease. Tuberculosis was defined as intrathoracic disease with positive sputum smears and/or cultures for *Mtb*. The study protocols were approved by the relevant human research ethics committees.

Three independent longitudinal South African progressor cohorts were analysed for the kinetics of the frequency changes in NK cells using deconvolution of gene expression datasets. These cohorts transitioned from: (1) an uninfected state to latency (GSE116014), for which samples were obtained from an *Mtb*-acquisition sub-cohort, selected from the larger ACS cohort^{5,16}, who were diagnosed as QFT-negative at multiple time points, six months apart, and then converted to QFT-positive, indicating a newly acquired *Mtb* infection (Extended Data Fig. 9); (2) latency to active disease, individuals enrolled in the ACS cohort described above, were assessed longitudinally every six months during a 2-year follow-up study¹⁵. Adolescents who developed active tuberculosis disease during this 2-year follow-up were included as ‘progressors’; and those who did not, were classified as LTBI ‘non-progressors’ (Supplementary Table 6). Active TB was defined as intrathoracic disease, with either two sputum smears that were positive for acid-fast bacilli or one positive sputum culture confirmed to be *M. tuberculosis* complex (mycobacterial growth indicator tube, BD BioSciences). Participants were excluded if they were known to be HIV-positive; (3) active disease to end-of-treatment, the Catalysis-TB cohort¹⁷.

Mass cytometry measurements and analysis. CyTOF experiments were performed as previously described³⁸. In brief, cryopreserved PBMCs were thawed and rested in complete RPMI with 10% FCS at 37 °C for 2 h at cell densities of approximately 10⁷ cells per ml. Cells from each sample were equally split into two parts and were either left untreated or stimulated for 4 h with 150 ng ml⁻¹ phorbol-12-myristate-13-acetate (PMA) and 1 μM ionomycin in the presence of brefeldin A and monensin (eBioscience). Cells (5 × 10⁶) were then stained (1 h; room temperature) with a mixture of metal-tagged antibodies (a complete list of antibodies and their catalogue numbers is provided in Supplementary Table 2). All antibodies were validated by the manufacturers for mass cytometry applications (as indicated on the manufacturer’s datasheet, available online) and were conjugated using MAXPAR reagents (Fluidigm Inc.). Cisplatin and iridium intercalators were used to identify live and dead cells. We used palladium barcoding (Fluidigm Inc.) according to the manufacturer’s instructions. Cells were washed twice with PBS, fixed in 1.6% paraformaldehyde (PFA) (Sigma-Aldrich; 1 h), washed again in ultrapure water and analysed using CyTOF mass cytometry on a CyTOF 2 instrument (Fluidigm). Intracellular phosphorylated-protein staining was carried out as previously described³⁹. In brief, cryopreserved PBMCs were thawed and rested as described above. Rested cells were incubated with cisplatin for 1 min and immediately quenched with four volumes of complete RPMI with 10% FCS, and rested again for 30 min at 37 °C. Subsequently, cells were split into five tubes; one was left untreated and the others were stimulated with (1) PMA and ionomycin, (2) IFNγ (50 ng ml⁻¹), (3) TNF (50 ng ml⁻¹) or (4) anti-CD3 (500 ng ml⁻¹) and anti-CD28 (2 μg ml⁻¹) for 15 min at 37 °C. The reaction was stopped by adding PBS with 2% PFA (incubated for 10 min; room temperature), followed by palladium barcoding as recommended by the manufacturer (Fluidigm). After barcoding, cell samples were washed and then combined for surface-marker staining (1 h; room temperature). Subsequently, cells were washed and permeabilized in MeOH at –80 °C overnight. The next day, cells were washed and incubated with the cocktail of antibodies to intracellular signalling proteins at room temperature for 1 h, followed by DNA staining as described above.

Cell events were acquired at approximately 500 events s⁻¹. In addition, we spiked each sample with internal metal-isotope bead standards for sample normalization using the CyTOF software (Fluidigm Inc.). Data processing and gating of dead cells and normalization beads was done on the Cytobank website (<http://www.cytobank.org>). To account for intra-run declines in mean marker intensity over time, we performed a within-sample-over-time normalization step by using a running window to adjust mean marker intensity throughout each individual run, such that the mean expression over time was equal to that measured at the beginning of the run. Data was debarcoded using Fluidigm’s Debarcoder tool. Data was arcsinh-transformed for analysis.

Analysis of CyTOF data. *Citrus* (cluster identification, characterization and regression). Cell subset abundance and functional marker expression in PBMCs from uninfected controls and subjects with LTBI were compared using the Citrus algorithm available from the Cytobank website. Citrus⁴⁰ uses regularized supervised learning algorithms to identify stratifying clusters (subsets) and cell response features. It is data driven and corrected for multivariate comparisons. In brief, Citrus analysis consists of the following steps. First, FCS files of normalized, ‘live cell/no beads’ samples were randomly sampled for *n* single-cell events. Second, collected single-cell events were pooled and iteratively hierarchically clustered based on similarity of expression of subsets of the measured channels. This procedure yielded overlapping clusters with the largest cluster encompassing all of the sampled events. Third, the pooled dataset was split back into its constitutive samples, and the relative abundance of cells in each cluster was computed, as well as the median expression of each functional marker in each cluster. Only clusters for which the abundance in one or more of the measured samples was greater than some lower-bound *P* values were considered for downstream differential analysis. Fourth, to determine differences in cell subset abundances or functional marker medians expression, we used the SAM algorithm in Citrus, which assesses FDR by permutations. For each set of analysis, we set *n* to 20,000 and the clustering threshold to 1% of total cells and performed the analysis iteratively such that 20,000 events from the entire dataset were chosen randomly for the multiple rounds of analysis. We also analysed the entire dataset in R. These analyses yielded qualitatively similar results.

Analyses of abundance from unstimulated and stimulated samples were done separately (see Extended Data Figs. 1, 2) because stimulation changed the expression levels of certain cell-surface markers. Manual inspection of Citrus output was used to identify the closest known gross-cell type. We characterized cell clusters using standard cell subset definitions: B cells (CD19⁺), CD8⁺ αβ T cells (CD3⁺TCRβ⁺CD8⁺), CD4⁺ αβ T cells (CD3⁺TCRβ⁺CD4⁺), γδ T cells (CD3⁺TCRγ⁺), monocytes (CD3[−]CD19[−]CD33⁺CD14⁺HLA-DR⁺), NK cells (CD3[−]CD19[−]CD14[−]HLA-DR[−]CD16⁺CD56^{bright/dim}). In all conditions, we report cluster abundance differences at a FDR < 1%.

viSNE analysis. Single-cell analysis using the dimensionality reduction technique viSNE reduces the multi-parametric data into two dimensions for visualization of similarity and heterogeneity across individual cells⁴¹. To account for different scales between parameters, the data was arcsinh transformed. viSNE analysis was performed on raw CyTOF data from the Cytobank database.

Manual gating. Manual gating was performed on the Cytobank website on normalized, debarcoded data files. A hierarchical gating strategy was used to identify live, single cells of the main PBMC populations (Extended Data Fig. 1a) and their subsets based on the expression of surface, cytokine or signalling molecules.

NK cell cytotoxicity measured by calcein-release assay. NK cells were enriched from PBMCs using the NK Cell Isolation Kit from Miltenyi Biotec. The NK cell cytotoxicity assay was carried out as described⁴² with some modifications. In brief, after cell counting, NK cells were mixed with calcein-acetoxymethyl (calcein-AM) labelled target K562 cells (which are susceptible to NK cell-mediated killing because of the lack of surface MHC class I expression), at an effector to target ratio of 2:1. For staining of the target cells, 2 mM of calcein-AM (Life Technologies) was added to the target cells (2 × 10⁶ per ml) and incubated at 37 °C for 30 min with periodic mixing. The target cells were washed, the enriched NK cells were added, and the mixture was incubated at 37 °C for 4 h. Maximum and spontaneous release controls were set up as three replicates using 1% Triton X-100 (final concentration) and plain medium, respectively. After the 4-h incubation, the cells were gently mixed to evenly distribute the released calcein in the supernatant and the plate was spun at 400g for 2 min to pellet the cells and any debris. For the calcein-release assay, 150 μl of the supernatant was collected and transferred to a flat black-bottom plate. The fluorescence was read using a FlexStation3 microplate reader (excitation/emission: 485/530 nm). The percentage of specific lysis was calculated using the formula: ((test release – spontaneous release)/(maximum release – spontaneous release)) × 100.

FcγR-mediated antibody-dependent cell mediated cytotoxicity. ADCC was carried out as previously described⁴³ with some modifications. In brief, P815 cells (a mouse leukaemia cell line) were stained with 0.25 μM carboxy-fluorescein succinimidyl ester (CFSE; Molecular Probes) and incubated with the 10 μg ml⁻¹

concentration of the P815-specific monoclonal antibody, 2.4G2. Coated and uncoated P815 cells were then cocultured with previously cryopreserved PBMC samples at an effector:target ratio of 10:1 in 10% FCS, penicillin, glutamine and streptomycin. The percentage of target cells that were killed through ADCC was monitored by flow cytometry staining using 7-amino-actinomycin D (7-AAD) viability staining solution (BioLegend). The percentage of cells killed by Fc γ R-mediated ADCC was obtained by subtracting the percentage 7-AAD⁺ CFSE-labelled uncoated target cells from the percentage of 7-AAD⁺ CFSE-labelled coated target cells. A minimum of 300,000 cells were analysed on a BD LSRII, and analysis was then performed using FlowJo (version 10.2) software.

Plasma protein quantification using a proximity extension assay. For analysis, 20 μ l of frozen (-80°C) plasma samples were thawed and sent to Olink Proteomics. In proximity extension assays, plasma proteins were dually recognized by pairs of antibodies coupled to a cDNA strand that ligates when brought into proximity to its target, extended by a polymerase and detected using a BioMark HD 96 \times 96 dynamic PCR array (Fluidigm). The quantification cycle (C_q) values from a DNA extension control are subtracted from the measured C_q value, an interpolate control is corrected for, and finally a correction factor is subtracted to yield a normalized protein expression value, which is \log_2 -transformed.

Cell-mixture deconvolution and cell percentage meta-analysis. This was carried out as previously described^{12,13}. In brief, publicly available gene expression datasets were collected, pre-processed using the MetaIntegrator R package⁴⁴ and annotated⁴⁵. Each microarray dataset was converted into a gene-expression matrix, whereas for GSE79362^{15,16}, which is an RNA-seq dataset, the total read counts per gene were first computed. A Hedge's g effect size was then computed to estimate changes in cell subset proportions. Effect sizes from all individual datasets were integrated into a summary effect size and significance was computed as previously described^{44,46}. To delineate cell trajectories over time, we computed a cumulative effect size score for each cell type in uninfected controls, LTBI, active TB disease, and end-of-treatment stages. We started by setting a reference effect size of 0 for healthy, *Mtb*-uninfected controls and then computed all following scores by adding the summary effect size value corresponding to case class for that specific comparison (LTBI, active TB, treated). We computed cumulative standard errors by assuming summary effect sizes to be normally distributed and independent of each other at each stage. Plots and statistics were generated using the R programming language.

Analysis of cell abundance using flow cytometry. For flow cytometry experiments, PBMCs were thawed and rested as described in 'Mass cytometry measurements and analysis'. Cells ($3\text{--}5 \times 10^6$) were incubated with antibodies for 30 min at 4°C , then washed with FACS staining buffer (PBS containing 1% bovine serum albumin and 0.05% sodium azide). The following monoclonal antibodies were used: Pacific Blue-conjugated anti-CD3 (BioLegend, 300417), PE-Cy7-conjugated anti-CD19 (BioLegend, 302216), APC-Cy7-conjugated anti-CD14 (BioLegend, 325620), APC-conjugated HLA-DR (BioLegend, 307610), PE-Dazzle-conjugated anti-CD16 (BioLegend, 302054) and Brilliant Violet 785-conjugated CD56 (BioLegend, 362550). The live/dead aqua-amine reactive dye was used for gating dead cells. All antibodies were validated by the manufacturers for flow cytometry application, as indicated on the manufacturer's website. Data were analysed using FlowJo version 10.2.

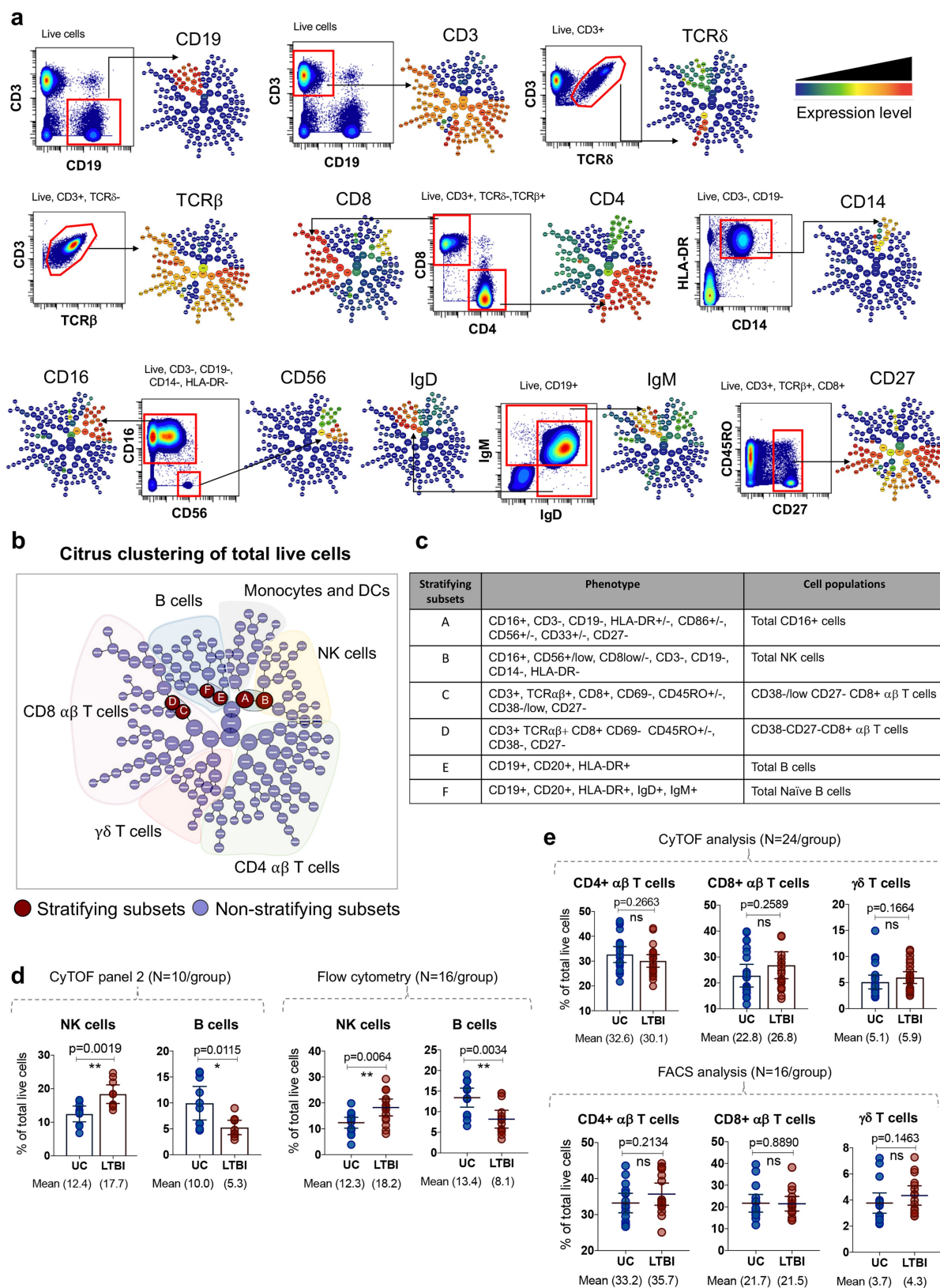
Statistical analysis. Analysis of CyTOF data are described in 'Citrus (cluster identification, characterization and regression)'. In all other experiments, significance levels were determined using Prism version 7 (GraphPad Software). Experiments were analysed using the Mann-Whitney U -test or one-way analysis of variance (ANOVA), as indicated for each experiment. The diagnostic performance of

NK cells to discriminate latent TB from active disease cases was evaluated using receiver operating characteristic (ROC) curve analysis, for which the true positive rate (sensitivity) is plotted as a function of the false-positive rate ($100 - \text{specificity}$). The area under the ROC curve is a measure of the probability that a classifier (for example, NK cell frequencies) will rank a randomly chosen positive instance (for example, active TB) higher than a randomly chosen negative one (for example, LTBI). ROC curves were plotted using Prism version 7.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The data that support the findings of this study are available from the corresponding authors upon reasonable request.

27. Berry, M. P. et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973–977 (2010).
28. Maertzdorf, J. et al. Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes Immun.* **12**, 15–22 (2011).
29. Maertzdorf, J. et al. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PLoS ONE* **6**, e26938 (2011).
30. Kaforou, M. et al. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Med.* **10**, e1001538 (2013).
31. Anderson, S. T. et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. *N. Engl. J. Med.* **370**, 1712–1723 (2014).
32. Bloom, C. I. et al. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PLoS ONE* **7**, e46191 (2012).
33. Verhagen, L. M. et al. A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children. *BMC Genomics* **14**, 74 (2013).
34. Cai, Y. et al. Increased complement C1q level marks active disease in human tuberculosis. *PLoS ONE* **9**, e92340 (2014).
35. Ottenhoff, T. H. et al. Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS ONE* **7**, e45839 (2012).
36. Tientcheu, L. D. et al. Differential transcriptomic and metabolic profiles of *M. africanum*- and *M. tuberculosis*-infected patients after, but not before, drug treatment. *Genes Immun.* **16**, 347–355 (2015).
37. Lee, S. W. et al. Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis. *BMC Bioinformatics* **17**, S3 (2016).
38. Leipold, M. D. & Maecker, H. T. Phenotyping of live human PBMC using CyTOFTM mass cytometry. *Bio Protoc.* **5**, e1382 (2015).
39. Fernandez, R. & Maecker, H. Cytokine-stimulated phosphoflow of PBMC using CyTOF mass cytometry. *Bio Protoc.* **5**, e1496 (2015).
40. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl Acad. Sci. USA* **111**, E2770–E2777 (2014).
41. Amir, E. D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
42. Somanchi, S. S., McCulley, K. J., Somanchi, A., Chan, L. L. & Lee, D. A. A novel method for assessment of natural killer cell cytotoxicity using image cytometry. *PLoS ONE* **10**, e0141074 (2015).
43. Salinas-Jazmín, N., Hisaki-Itaya, E. & Velasco-Velázquez, M. A. A flow cytometry-based assay for the evaluation of antibody-dependent cell-mediated cytotoxicity (ADCC) in cancer cells. *Methods Mol. Biol.* **1165**, 241–252 (2014).
44. Haynes, W. A. et al. Empowering multi-cohort gene expression analysis to increase reproducibility. *Pac. Symp. Biocomput.* **22**, 144–153 (2017).
45. Sweeney, T. E., Braviak, L., Tato, C. M. & Khatri, P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *Lancet Respir. Med.* **4**, 213–224 (2016).
46. Khatri, P. et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. *J. Exp. Med.* **210**, 2205–2221 (2013).

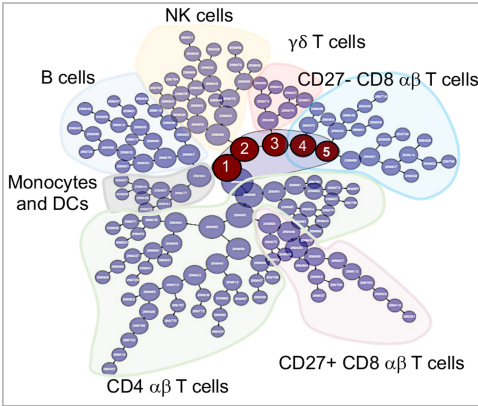


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Broad alterations of peripheral immune cell distributions in LTBI. PBMCs from 14 latently infected and 14 uninfected participants of a South African adolescent cohort were characterized using CyTOF with antibody panel 1 (Supplementary Table 2), followed by Citrus analysis and clustering. This unsupervised hierarchical clustering analysis produced a branching structure (dendrogram) that allowed the grouping of total live cells into known immune cell compartments (contoured). Cell clusters are represented as nodes (circles) in this Citrus-derived circular dendrogram, which delineates lineage relationships that were identified from the data. Cluster granularity (that is, cell subset specificity) increases from the centre of the diagram to the periphery. **a**, Annotation of cluster hierarchy plots based on surface marker expression. The expression intensity of each marker used for cell population characterization is overlaid per cluster on the Citrus circular dendrogram and is indicated, independently for each marker, by the coloured gradient for which the range corresponds to the arcsinh-transformed expression of the median marker expression measured across all Citrus clusters. For each marker, we also provide a dot plot graph demonstrating the marker labelling in the manually gated indicated

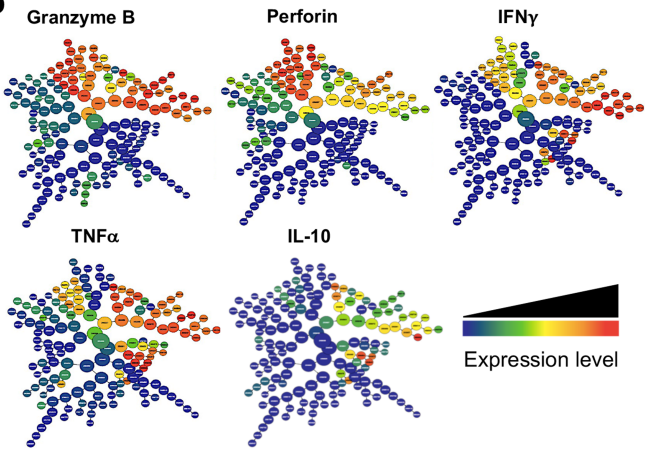
population. **b**, Citrus plots showing, based on cell-surface protein expression, clusters (in red, designated A–F) that exhibit significantly different abundances (SAM analysis with FDR < 1%) between the uninfected and latently infected individuals. Individual cell clusters are mapped to well-established, gross-cell types: B cells (CD19⁺), CD8⁺ $\alpha\beta$ T cells (CD3⁺TCR β ⁺CD8⁺), CD4⁺ $\alpha\beta$ T cells (CD3⁺TCR β ⁺CD4⁺), $\gamma\delta$ T cells (CD3⁺TCR δ ⁺), monocytes (CD3[−]CD19[−]CD33⁺CD14⁺HLA-DR⁺), NK cells (CD3[−]CD19[−]CD14[−]HLA-DR[−]CD16⁺CD56^{bright/dim}), identifiable by annotated shaded background groupings. **c**, The phenotype and the composition of cells in each of the stratifying cell subsets (A–F), identified by Citrus analysis. **d**, Percentages of NK cells and B cells determined by manual gating of 20 additional samples using CyTOF antibody panel 2 (left; Supplementary Table 2) and 32 samples using flow cytometry (right). **e**, Percentages of CD4⁺ $\alpha\beta$ T cells, CD8⁺ $\alpha\beta$ T cells and $\gamma\delta$ T cells in uninfected controls and latently infected individuals, analysed by CyTOF ($n = 24$ per group; top) and flow cytometry ($n = 16$ per group; bottom). Throughout, P values were derived using a Mann–Whitney U -test. Mean and error bars representing the 95% confidence intervals are shown for each comparison.

a Citrus clustering of total live cells (stimulated)



● Stratifying subsets ● Non-stratifying subsets

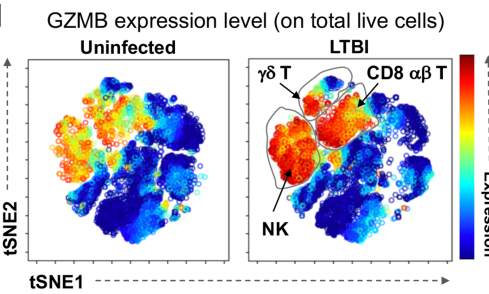
b



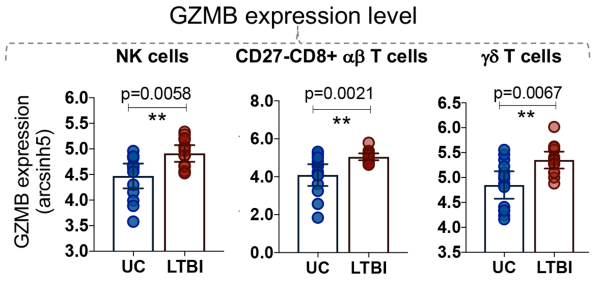
c

Stratifying subsets	Phenotype	Cell populations
1	GZMB+PRF+	NK cells, Monocytes, DCs, B cells, CD27-CD8 αβ and γδ T cells
2	GZMB+PRF+IFNγ+TNFα+	NK cells, CD27- CD8 αβ T and γδ T cells
3	GZMB+PRF+IFNγ+TNFα ^{hi}	CD27-CD8 αβ T and γδ T cells
4	GZMB+PRF+IFNγ+TNFα ^{hi}	CD27-CD8 αβ T cells
5	GZMB+PRF+IFNγ+TNFα ^{hi} IL-10 ^{low}	CD27-CD8 αβ T cells

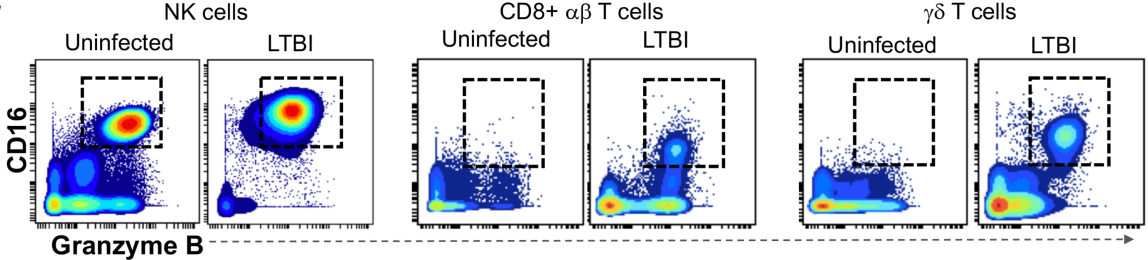
d



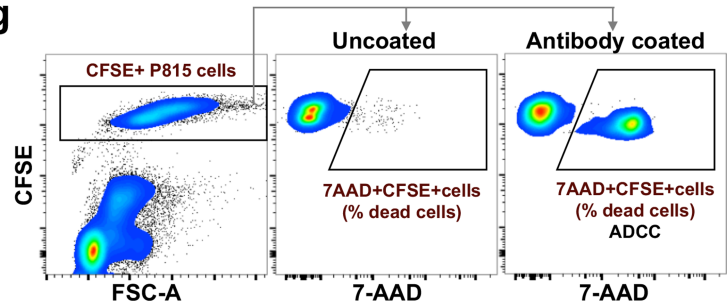
e



f



g

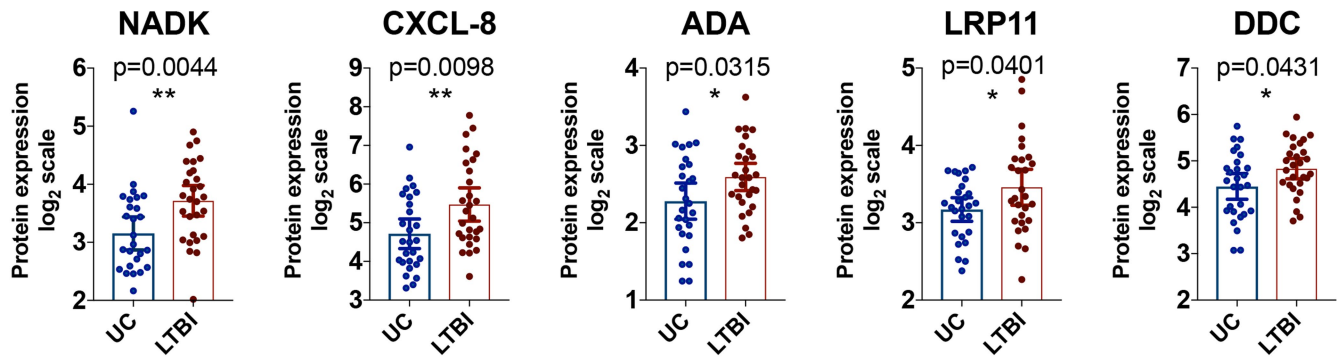
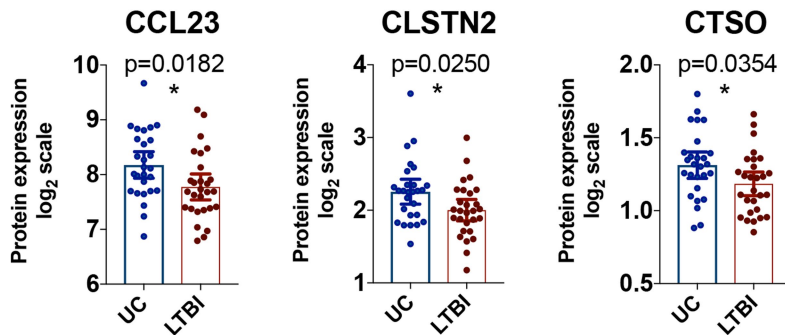


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Enhanced effector function response in LTBI.

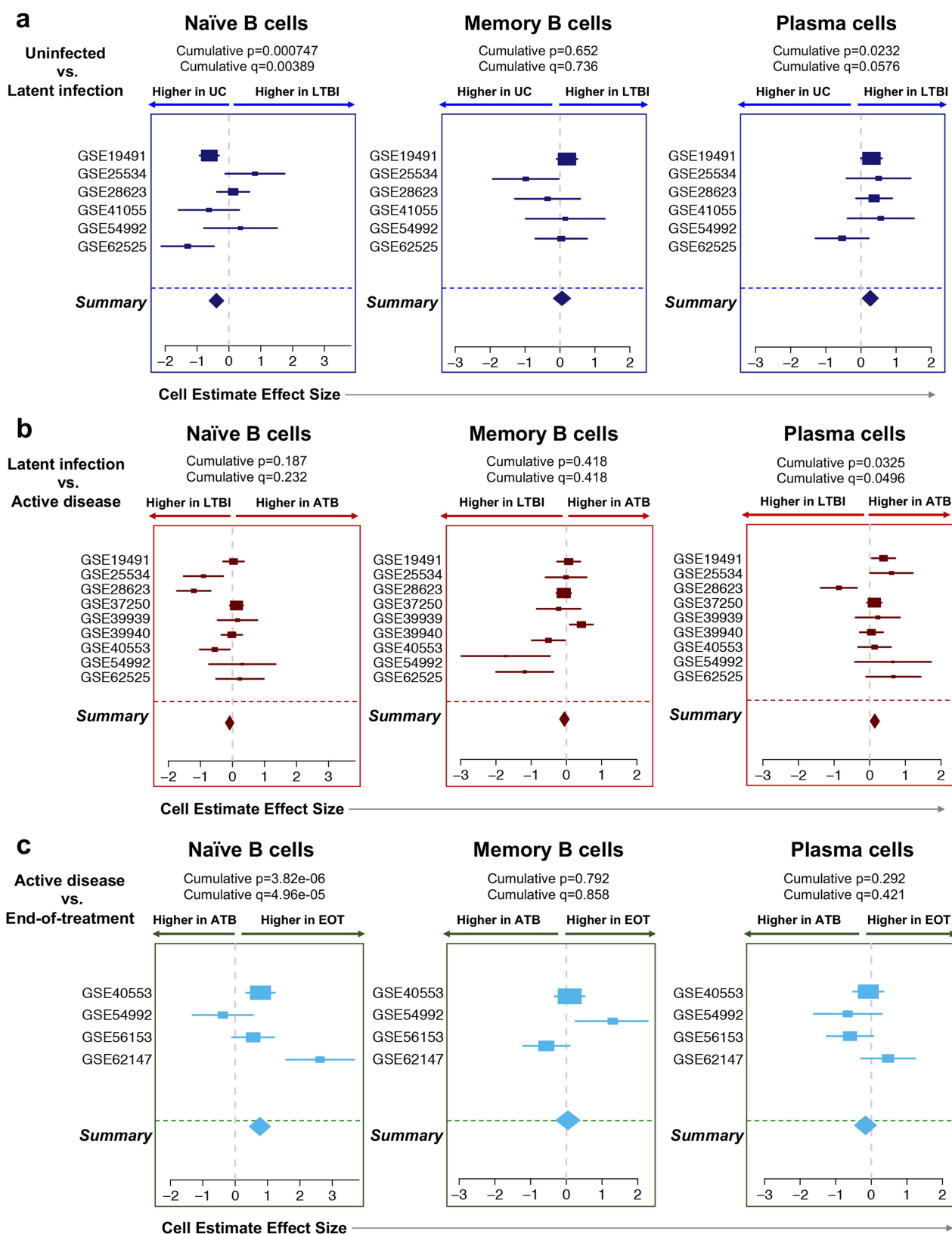
a, Cell subsets, shown as red nodes in a Citrus-derived circular dendrogram and designated as 1–5, were identified as significantly different in abundance (SAM analysis at $FDR < 1\%$) based on CyTOF analysis of effector and cell-surface molecule expression on PBMCs (antibody panel 1, Supplementary Table 2) from uninfected controls and individuals with LTBI ($n = 14$ per group) after 4-h PMA and ionomycin stimulation. Mapping of individual cell clusters to established, gross-cell types are identified by annotated shaded background groupings. **b**, Expression intensity of selected effector molecules is indicated by the coloured gradient for which the range corresponds to the arcsinh-transformed expression of the median marker expression measured across all Citrus clusters. **c**, Effector molecule expression and the composition of cells in each of the stratifying cell clusters (1–5), identified by Citrus

analysis. **d**, viSNE analysis of GZMB expression level in immune-cell subsets, representative of 14 uninfected and 14 individuals with LTBI (the colour gradient corresponds to the arcsinh-transformed expression level). **e**, Quantification of intracellular GZMB expression level in NK cells, $CD8^+ \alpha\beta$ T cells and $\gamma\delta$ T cells in uninfected controls and individuals with LTBI ($n = 14$ per group). P values were derived using a Mann–Whitney U -test. Mean and error bars representing the 95% confidence intervals are shown for each comparison. **f**, Dot plots from CyTOF analysis of $CD16^+ GZMB^{high}$ cells within each lymphocyte subset, representative of 14 uninfected controls and 14 individuals with LTBI. **g**, Gating strategy for ADCC. ADCC was measured using NK-resistant P815 cells, which were either coated with antibody (2.4G2) or left uncoated (control), and labelled with the intracellular dye CFSE, followed by the DNA dye 7AAD. $CFSE^+ 7AAD^+$ cells were defined as dead target cells.

a Plasma proteins increased in latent TB infection**b** Plasma proteins decreased in latent TB infection

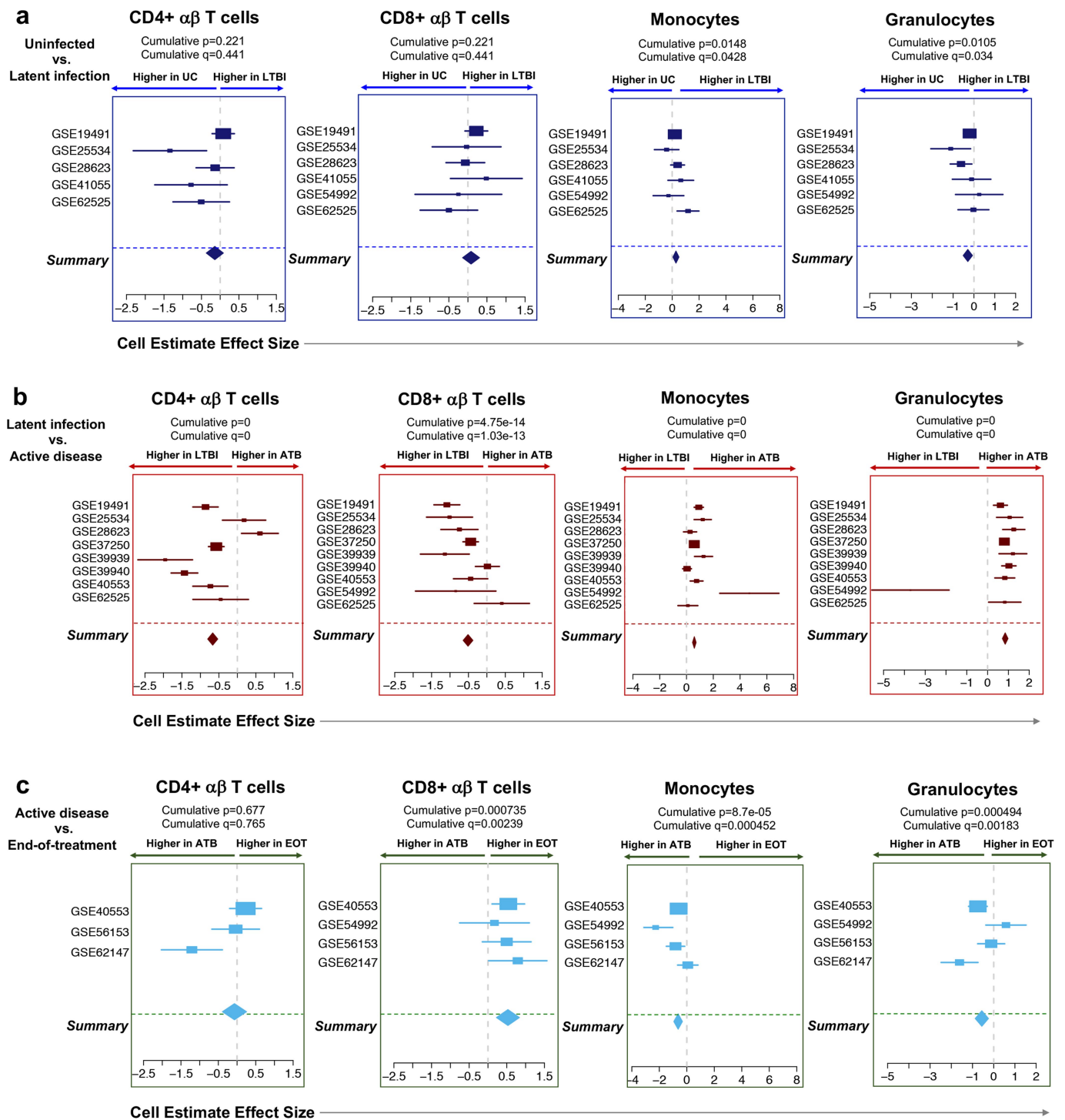
Extended Data Fig. 3 | Alterations in plasma protein levels in LTBI. The relative levels of plasma proteins (Supplementary Table 3), shown on a log₂ scale, between uninfected controls and individuals with LTBI ($n = 27$ per group). Plasma proteins that were present at significantly higher levels (a) and significantly lower levels (b) in individuals with LTBI. Plasma protein

quantification was performed using the proximity extension assay. *P* values were derived using an unpaired two-tailed Student's *t*-test. Mean and error bars representing the 95% confidence intervals are shown for each comparison.



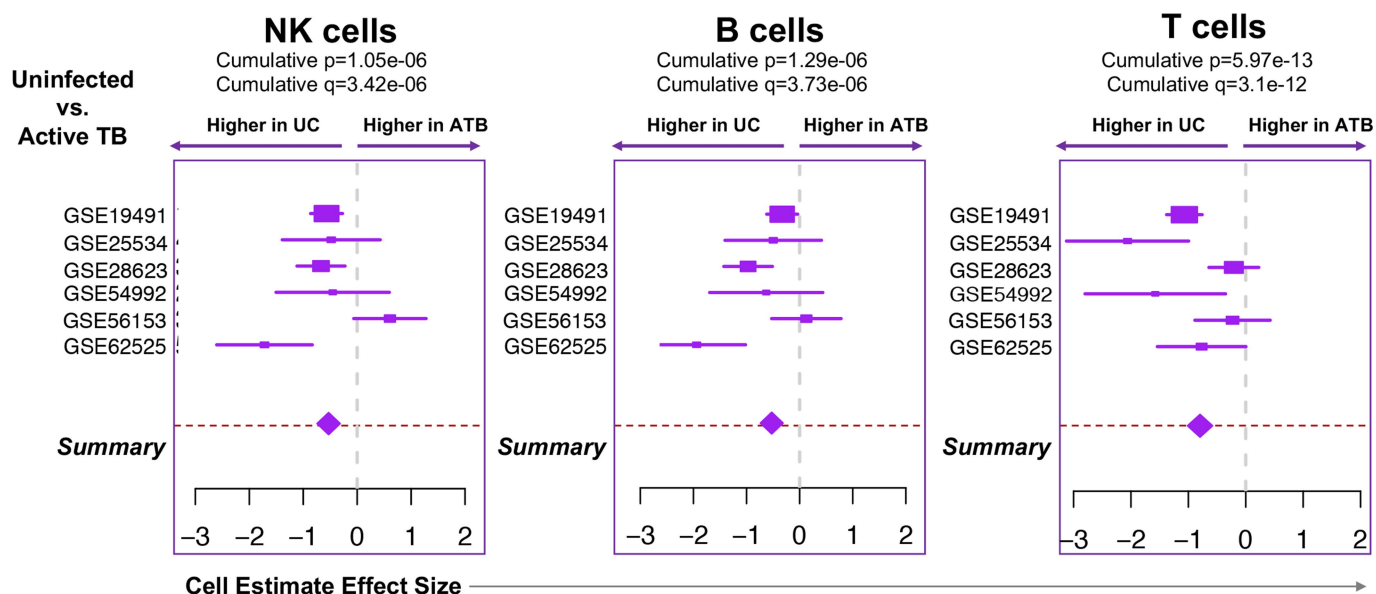
Extended Data Fig. 4 | Changes in frequencies of peripheral B cell subsets in LTBI, active TB and after treatment. Forest plots for estimated frequencies of B cell subsets: naïve B cells, memory B cells and plasma cells. **a**, Comparison between the uninfected state ($n = 189$) and LTBI ($n = 145$). **b**, Comparison between LTBI ($n = 409$) and active TB ($n = 543$). **c**, Comparison between active TB ($n = 76$) and end-of-treatment ($n = 97$). Cohort GSE identifiers are listed on the left. In the plots, boxes represent the standardized mean difference in estimated

cellular proportions in a cohort between two comparison groups. The size of the box is proportional to the sample size of a given cohort. Lines indicate the 95% confidence interval of the corresponding effect sizes. Diamonds indicate the summary effect size (Summary), obtained by integrating the effect sizes from individual cohorts. The width of the diamond corresponds to its 95% confidence interval. The P values and q values for the summary effect sizes are shown above each plot.



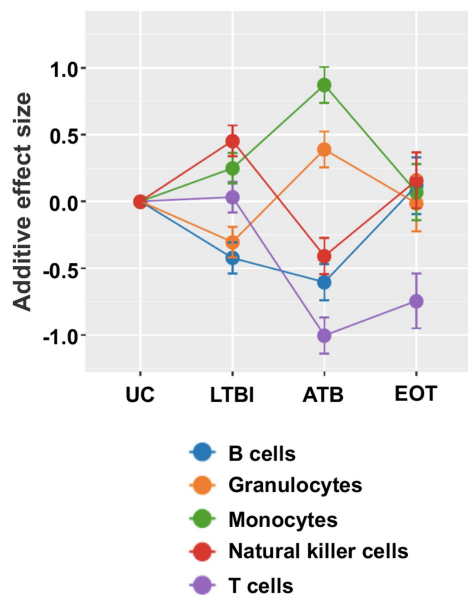
Extended Data Fig. 5 | Changes in frequencies of peripheral T cell subsets, monocytes and granulocytes in LTBI, active TB and after treatment. Forest plots for estimated frequencies of CD4⁺ αβ T cells, CD8⁺ αβ T cells, monocytes and granulocytes. **a**, Comparison between the uninfected state ($n = 189$) and LTBI ($n = 145$). **b**, Comparison between LTBI ($n = 409$) and active TB ($n = 543$). **c**, Comparison between active TB ($n = 76$) and end-of-treatment ($n = 97$). Boxes represent the standardized mean difference in estimated cellular proportions in a cohort between

two comparison groups. The size of the box is proportional to the sample size of a given cohort. Lines indicate the 95% confidence interval of the corresponding effect sizes. Diamonds indicate the summary effect size (Summary), obtained by integrating the effect sizes from individual cohorts. The width of the diamond corresponds to its 95% confidence interval. The P values and q values for the summary effect sizes are shown above each plot.

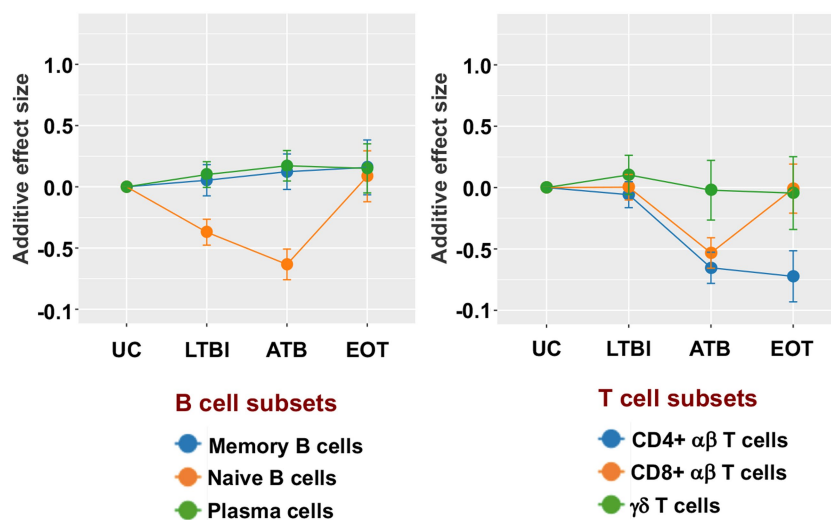


Extended Data Fig. 6 | Comparison of the frequencies of peripheral NK cells, B cells and T cells between uninfected controls and patients with active TB. Forest plots comparing changes in the levels of NK cells, B cells and T cells between uninfected individuals ($n = 191$) and patients with active TB ($n = 178$). Boxes represent the standardized mean difference in estimated cellular proportions in a cohort between two comparison

groups. The size of the box is proportional to the sample size of a given cohort. Lines indicate the 95% confidence interval of the corresponding effect sizes. Diamonds indicate the summary effect size (Summary), obtained by integrating the effect sizes from individual cohorts. The width of the diamond corresponds to its 95% confidence interval. The P values and q values for the summary effect sizes are shown above each plot.

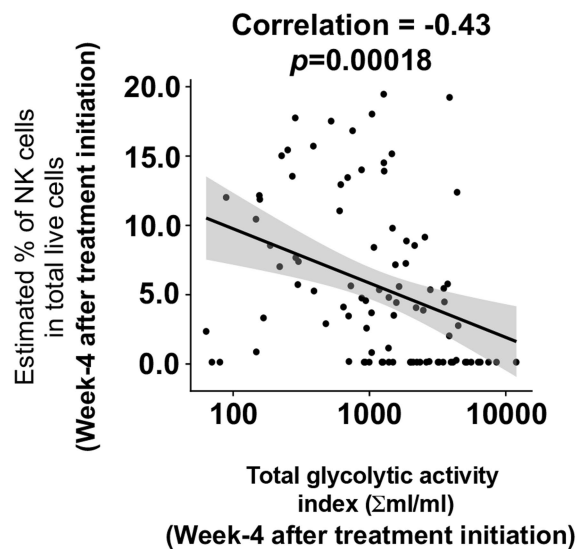
a Immune cell trajectories from acquisition of *Mtb* infection to treatment completion

Extended Data Fig. 7 | Trajectories of different immune cell populations from the acquisition of *Mtb* infection to end-of-treatment. Changes in the frequency distribution patterns of different peripheral leukocyte populations (a) and B and T cell subpopulations (b) at the

b B and T cell subset trajectories from acquisition of *Mtb* infection to treatment completion

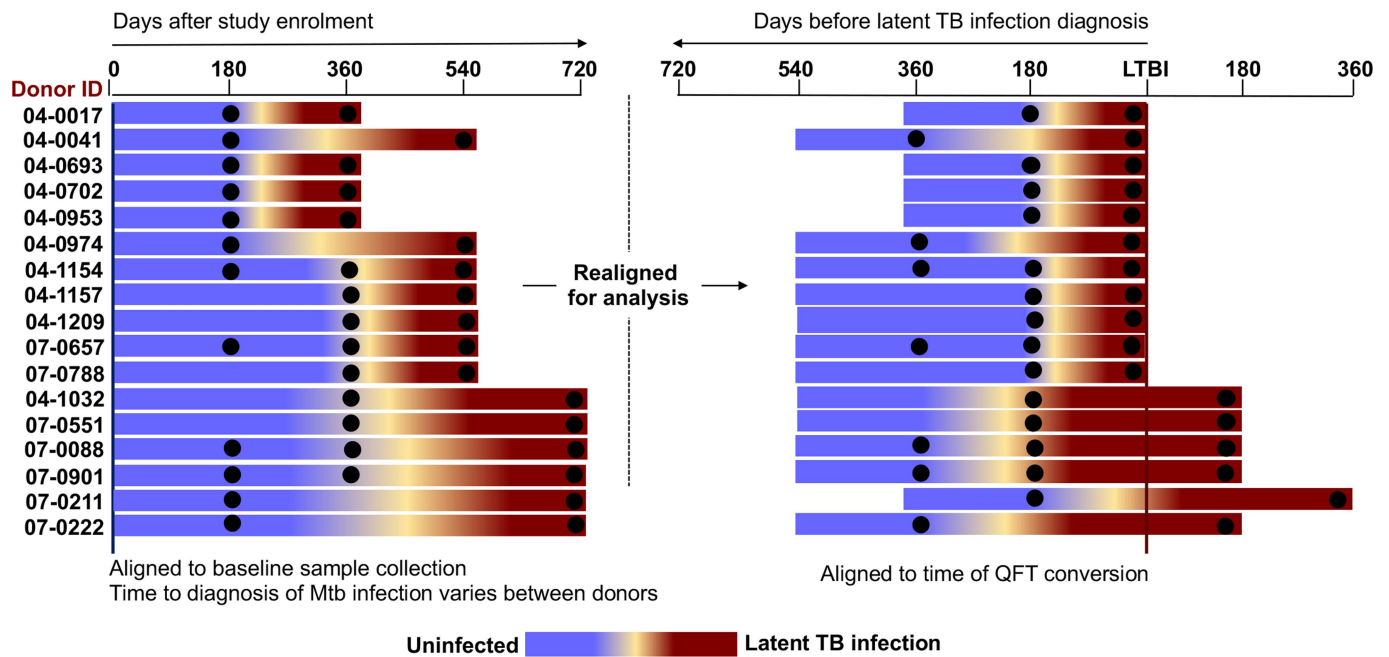
different stages of infection. Lines indicate cumulative effect size scores starting from a healthy baseline level up to treatment of active TB disease. Error bars indicate corresponding standard errors.

Correlations between peripheral NK cell % (in total live cells) and pulmonary inflammatory burden



Extended Data Fig. 8 | Correlation between peripheral NK cell percentage and lung inflammation. Correlation plot showing the relationship between estimated peripheral NK cell frequencies in patients with active TB at week 4 after treatment initiation and total glycolytic

activity index (TGAI) of the lung measured by PET-CT imaging at the corresponding time point. The line represents the best fit and the shaded area the 95% confidence interval. NK cell frequencies were determined by deconvolution.



Extended Data Fig. 9 | Synchronization of the adolescent cohort who underwent QuantiFERON conversion following *Mtb* acquisition .

To identify changes in peripheral NK cell frequencies after acquisition of *Mtb* infection by cell-mixture deconvolution analysis, the timescale of the gene expression dataset (GSE116014) was realigned according to the time of first infection diagnosis instead of study enrolment, allowing the identification of gene-expression profiles obtained before infection diagnosis. Each individual is represented by a horizontal bar. The length

of the bar represents the number of days between study enrolment and diagnosis with *Mtb* infection. During follow-up, each individual transitioned from an uninfected state (blue) to infected state (brown), that is, underwent QFT conversion. The black circles represent time points for which gene-expression data were available. Pre-infection (Pre) data (180–360 days) were compared to data obtained at the time of infection diagnosis or the nearest time point after diagnosis (Post) (0–360 days).

Allergic inflammatory memory in human respiratory epithelial progenitor cells

Jose Ordovas-Montanes^{1,2,3,4,5,6,14}, Daniel F. Dwyer^{7,8,14}, Sarah K. Nyquist^{1,2,3,4,5,9,10}, Kathleen M. Buchheit^{7,8}, Marko Vukovic^{1,2,3,4,5}, Chaarushena Deb^{1,2,3,4,5}, Marc H. Wadsworth II^{1,2,3,4,5}, Travis K. Hughes^{1,2,3,4,5}, Samuel W. Kazer^{1,2,3,4,5}, Eri Yoshimoto^{7,8}, Katherine N. Cahill^{7,8}, Neil Bhattacharyya^{8,11}, Howard R. Katz^{7,8}, Bonnie Berger^{10,12,13}, Tanya M. Laidlaw^{7,8}, Joshua A. Boyce^{7,8}, Nora A. Barrett^{7,8,15*} & Alex K. Shalek^{1,2,3,4,5,13,15*}

Barrier tissue dysfunction is a fundamental feature of chronic human inflammatory diseases¹. Specialized subsets of epithelial cells—including secretory and ciliated cells—differentiate from basal stem cells to collectively protect the upper airway^{2–4}. Allergic inflammation can develop from persistent activation⁵ of type 2 immunity⁶ in the upper airway, resulting in chronic rhinosinusitis, which ranges in severity from rhinitis to severe nasal polyps⁷. Basal cell hyperplasia is a hallmark of severe disease^{7–9}, but it is not known how these progenitor cells^{2,10,11} contribute to clinical presentation and barrier tissue dysfunction in humans. Here we profile primary human surgical chronic rhinosinusitis samples (18,036 cells, $n = 12$) that span the disease spectrum using Seq-Well for massively parallel single-cell RNA sequencing¹², report transcriptomes for human respiratory epithelial, immune and stromal cell types and subsets from a type 2 inflammatory disease, and map key mediators. By comparison with nasal scrapings (18,704 cells, $n = 9$), we define signatures of core, healthy, inflamed and polyp secretory cells. We reveal marked differences between the epithelial compartments of the non-polyp and polyp cellular ecosystems, identifying and validating a global reduction in cellular diversity of polyps characterized by basal cell hyperplasia, concomitant decreases in glandular cells, and phenotypic shifts in secretory cell antimicrobial expression. We detect an aberrant basal progenitor differentiation trajectory in polyps, and propose cell-intrinsic¹³, epigenetic^{14,15} and extrinsic factors^{11,16,17} that lock polyp basal cells into this uncommitted state. Finally, we functionally demonstrate that ex vivo cultured basal cells retain intrinsic memory of IL-4/IL-13 exposure, and test the potential for clinical blockade of the IL-4 receptor α -subunit to modify basal and secretory cell states in vivo. Overall, we find that reduced epithelial diversity stemming from functional shifts in basal cells is a key characteristic of type 2 immune-mediated barrier tissue dysfunction. Our results demonstrate that epithelial stem cells may contribute to the persistence of human disease by serving as repositories for allergic memories.

The type 2 immunity (T2I) module⁶ regulates homeostatic processes¹⁸ (metabolism), host defence¹⁹ (against parasites, venoms, allergens and toxins), and inflammatory tissue repair¹¹. However, this module may become self-reinforcing in allergic inflammation, leading to substantial alterations in gross tissue architecture²⁰ as observed in polyps⁷. To investigate how the overall tissue cellular ecosystem shifts in composition and states during chronic respiratory T2I disease in humans, we used Seq-Well for single-cell RNA sequencing (scRNA-seq)¹² to profile the ethmoid sinus (EthSin) of patients spanning the

chronic rhinosinusitis (CRS) spectrum (Fig. 1a, Supplementary Table 1, Methods; Supplementary Discussion I; $n = 12$ samples: 6 non-polyp, 6 polyp). Deconstructing these tissues into their component cells provides a unique lens into the cellular ecosystem of human T2I, helping us to: 1. characterize each major cell type without the biases that are typically introduced by pre-selection of markers; 2. evaluate cell types and/or states with disease-associated transcriptional differences; and, 3. reconstruct tissue-level dynamics.

We derived a unified cells-by-genes expression matrix (18,036 cells) and performed dimensionality reduction and graph-based clustering (Fig. 1a, Extended Data Fig. 1a, b, Supplementary Table 2; Methods). Using complete lists of cluster-specific genes to identify epithelial², stromal^{7,20} and immune cells^{4,6}, we recovered a reproducible distribution of cell types within patient groups (Fig. 1b, c, Extended Data Figs. 1c–e, 2a–e, Supplementary Table 3; Methods; Supplementary Discussion II). We highlight the major cell types recovered (with hallmark expressed genes in parentheses): basal (*KRT5*) and apical (*KRT8*) cells—which orient the pseudostratified epithelial division, further specialized ciliated (*FOXJ1*) and glandular³ (*LTF*) cells, and supportive endothelial cells (*DARC*, also known as *ACKR1*), fibroblasts (*COL1A2*), plasma cells (*CD79A*), myeloid cells (*HLA-DRA*), T cells (*TRBC2*) and mast cells (*TPSAB1*) (Extended Data Fig. 1e). For each cell type, sub-clustering revealed further, potentially meaningful heterogeneity, providing a useful reference atlas for studying human inflammatory diseases of barrier tissues (Extended Data Fig. 3a–c; Supplementary Discussion III).

Next, we charted the cell-of-origin for chemokines and lipid mediators, which aid in the recruitment and positioning of lymphoid and myeloid cells in tissues during T2I²¹ (Extended Data Fig. 4a; Supplementary Discussion IV). For example, we found mast cells specifically enriched for *HPGDS* and *PTGS2*, suggesting that they may be a dominant source of prostaglandin D₂, which is implicated in activation of T helper 2 (Th2) cells⁴. Alongside these mediators, the production of instructive first-order cytokines primes recruitment and activation of effector mechanisms. In particular, IL-25, IL-33 and thymic stromal lymphopoietin (TSLP) are broadly regarded as epithelial-derived cytokines^{4,5,16,20,22}, yet little is known about which cells express them in human disease. *TSLP* expression was uniquely restricted to basal cells, suggesting a link between increased basal cell numbers in disease and activation of effector cells (Fig. 1d, Extended Data Figs. 3a, 4b, c; Supplementary Discussion IV).

Expression of second-order effector cytokines was identified in a subset of CD4⁺ T cells expressing *IL4*, *IL5*, *IL13* and *HPGDS*, fitting the profile of allergen-specific Th2A cells²³ (Fig. 1d, Extended Data Fig. 4c–e;

¹Institute for Medical Engineering and Science (IMES), Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA. ⁴Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁵Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. ⁶Division of Infectious Diseases and Division of Gastroenterology, Boston Children's Hospital, Boston, MA, USA. ⁷Jeff and Penny Vinik Center for Allergic Disease Research, Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, MA, USA. ⁸Department of Medicine, Harvard Medical School, Boston, MA, USA. ⁹Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹⁰Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹¹Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA. ¹²Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹³Harvard-MIT Division of Health Sciences & Technology, Cambridge, MA, USA. ¹⁴These authors contributed equally: Jose Ordovas-Montanes, Daniel F. Dwyer. ¹⁵These authors jointly supervised this work: Nora A. Barrett, Alex K. Shalek. *e-mail: nbarrett@bwh.harvard.edu; shalek@mit.edu

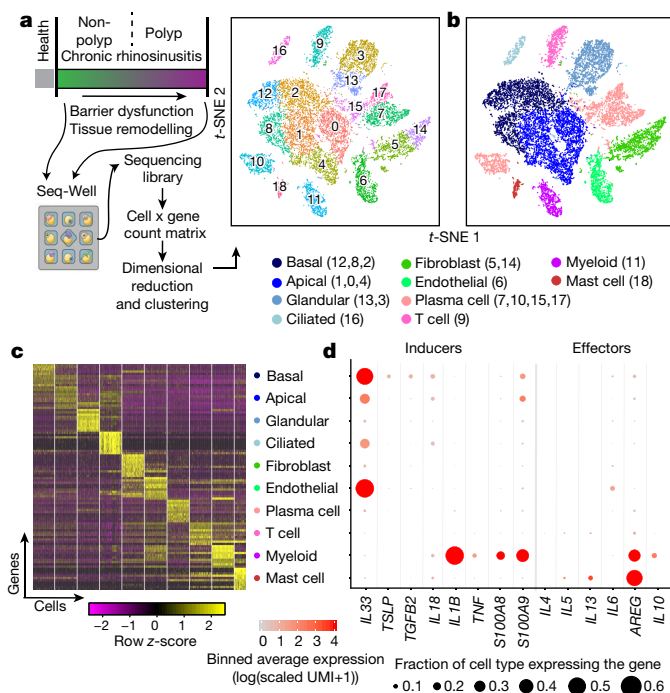


Fig. 1 | Mapping the T2I inflamed human sinus cellular ecosystem by scRNA-seq. **a**, Clinical disease spectrum ($n = 12$ samples) and experimental workflow leading to a t -distributed stochastic neighbour embedding (t -SNE) plot displaying 18,036 single cells, coloured by shared nearest neighbour (SNN) clusters (**a**) and cell types (**b**) (receiver operating characteristic (ROC) test; Supplementary Table 3; Methods) from respiratory tissue. **c**, Heat map of top 10 marker genes by ROC test (**c**) (area under curve (AUC) > 0.73) for indicated cell types; maximum 500 cells plotted per type (Extended Data Fig. 2a annotated; Supplementary Table 3 full gene list) and dot plot of T2I mediators mapped onto cell types (**d**) across all samples (Extended Data Fig. 4 shows this by disease state).

Supplementary Discussion V). Additionally, substantial numbers of mast cells expressed *IL5* and *IL13*, and, along with myeloid cells, were the main expressers of the tissue-reparative cytokine *AREG*²². Notably, patients with or without polyps showed consistent cells-of-origin for T2I-related chemokines, lipids and cytokines, with the exception of select mediators (Extended Data Fig. 4a, b; Supplementary Discussion IV). Expression of several genes implicated in allergic diseases²⁴ by genome-wide association studies (GWAS) was restricted to specific cell types. We therefore mapped the expression of candidate risk genes, including *GATA2*, *IL1RL1* (which encodes the IL-33 receptor ST2 subunit), *CDHR3*, *KIF3A*, *TMEM232* and *MYC* (Extended Data Fig. 4f; Supplementary Discussion VI). Cellular maps of tissues commonly affected by inflammatory disease should help to provide mechanistic insights into genotype–phenotype interactions.

We further analysed the epithelial clusters (Fig. 2a, Extended Data Fig. 5a–c), providing single-cell human transcriptomes²⁵ for basal, secretory, glandular and ciliated cell types from a T2I ecosystem (Fig. 2a, b, Extended Data Fig. 5, Supplementary Table 3). Analysis of epithelial marker genes identified conserved transcriptional programs in basal (three clusters), differentiating/secretory (three clusters), glandular (two clusters) and ciliated (one cluster) cell types^{2,3} (Fig. 2a, b, Extended Data Fig. 5a–d, Supplementary Table 3; Supplementary Discussion VII).

On the basis of our observation of striking polyposis-related differences across clusters within cell types (Fig. 2c, Extended Data Fig. 5e; Supplementary Discussion VIII), we quantified the numerical over-representation of cells from the non-polyp and polyp ecosystems within each cluster and type. The clusters comprising basal, differentiating/secretory, and glandular cells showed the most significant links to the disease state (P values by Fisher's least-significant difference

test; Fig. 2c, Supplementary Table 3). We compared transcriptomes of differentiating/secretory cells³ (containing *KRT8*-expressing secretory and apical goblet cells), noting that secretory cells from polyps appear to supplant antimicrobial function with tissue repair (Fig. 2d, Supplementary Table 3; Supplementary Discussion VIII).

Of note, we observed expression of *MUC5B* within glandular mucus cells (cluster 13), whereas *MUC5AC* was expressed in a distinct subset of secretory goblet cells co-expressing *SCGB1A1* and *FOXA3* (Fig. 2b, Extended Data Fig. 5f, g; Supplementary Discussion IX). This suggests that the goblet cell program is overlaid on a secretory cell base². We also assessed glandular heterogeneity, identifying five discrete subsets with variegated antimicrobial expression³ (Fig. 2a, Extended Data Fig. 6a, b, Supplementary Table 3; Supplementary Discussion IX). This compartmentalization may represent a mechanism for regulated secretion, with imbalances in cell types or states affecting innate host defence.

To contextualize shifts associated with disease state, we turned to sinonasal scrapings as a method of sampling healthy apical cells through Seq-Well (Extended Data Fig. 6c, d, Supplementary Tables 3, 6; 18,704 additional cells: $n = 3$ healthy inferior turbinate (InfTurb), $n = 4$ polyp-patient InfTurb, $n = 2$ EthSin-polyp directly). We recovered immune cells and differentiating/secretory and ciliated epithelial cells from the InfTurb of patients with polyposis and healthy controls, but basal cells were found only in polyp scrapings (Extended Data Fig. 6d–f, Supplementary Table 3; Supplementary Discussion X). By combining all epithelial cells from the surgical resections with scrapings (Fig. 2a–c, e), we identified a conserved core secretory gene set that was present in all sites sampled, as well as healthy, CRS-InfTurb, CRS-EthSin-non-polyp and CRS-EthSin-polyp specific gene signatures. Overall, we note a shift from IFN- α /IFN- γ -induced genes to IL-4/IL-13-induced genes with increasing disease severity (Fig. 2e–g, Supplementary Table 3; Supplementary Discussion XI). Secretory cells from involved CRS-EthSin tissue differ markedly from those of the InfTurb, and secretory cells in non-polyp and polyp EthSin reach distinct states in which altered functionality may be linked to severity of disease.

As specialized epithelial cell types arise from basal progenitors^{2,10}, we formally examined their distribution in each sample (Fig. 3a, Extended Data Fig. 7a). Our data indicate a significant loss of epithelial ecological diversity in nasal polyps by Simpson's index (see Methods), largely driven by glandular and ciliated cell depletion, and an enrichment in basal cells (Fig. 3a, b, Extended Data Fig. 7a–d; Supplementary Discussion XII). This altered diversity tracked closely with rank-ordered pathology of patient tissue samples, which correlated positively with basal cell frequency ($r = 0.6252$) and negatively with epithelial diversity ($r = -0.6824$; Extended Data Fig. 7e). We hypothesize that alterations in the immune compartment in polyps may represent an overcorrection in attempting to balance epithelial shifts (Extended Data Fig. 7f; Supplementary Discussion XII).

To confirm our findings on epithelial cell types, we applied complementary approaches. Using flow cytometry¹⁰, we demonstrated that the frequency of basal cells significantly increased in polyps at the expense of differentiated epithelial cells in 13 additional patients (Fig. 3c, Extended Data Fig. 7g, h). Using histology (which, unlike scRNA-seq or flow cytometry, is not subject to dissociation-induced artefacts), we confirmed⁸ a significant increase of p63⁺ cells per 1,000 μm^2 of epithelial area and a striking loss of glands in polyps (Fig. 3d, e, Extended Data Fig. 7i, j). We also used marker genes for specialized lineages to deconvolve bulk EthSin-tissue RNA sequencing (RNA-seq) from an additional cohort of 27 individuals. We identified four patient clusters and confirmed glandular enrichment in non-polyps, and shifts in secretory cell states and the progressive acquisition of basal-associated transcripts in polyps (Fig. 2d, f, 3f, g, h, Supplementary Tables 1, 3; Supplementary Discussion XIII). Finally, we validated these findings with publicly-available RNA-seq datasets containing normal human sinus tissue and polyps (Extended Data Fig. 7k, l; Supplementary Discussion XIII).

To investigate mechanisms that might account for the reduced epithelial diversity in polyps, we compared the transcriptomes of

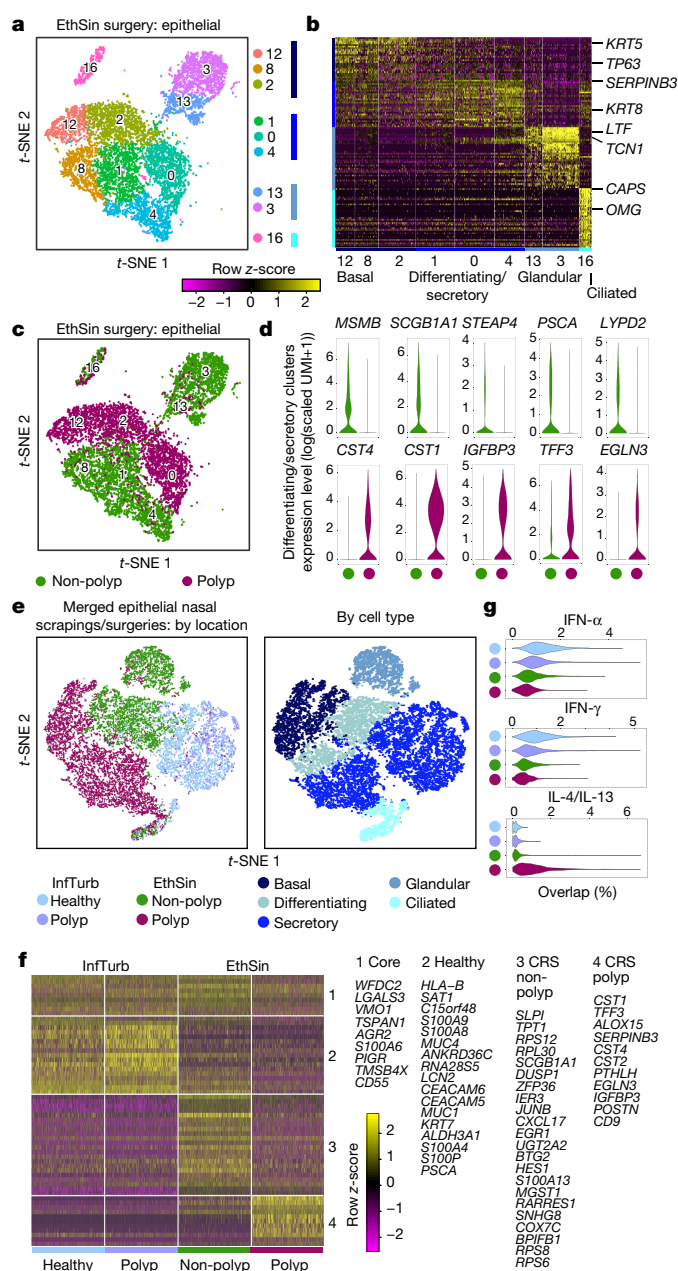


Fig. 2 | Single-cell transcriptomes of epithelial cells in T2I highlight shifts in secretory cell states across health and disease. **a, b,** *t*-SNE plot of 10,274 epithelial cells ($n = 12$ samples), coloured by SNN clusters (Fig. 1; re-clustered in Extended Data Fig. 6) with blue colour bars representing cell types determined according to Extended Data Fig. 5 (**a**) and heat map of marker genes by ROC (**b**) ($AUC > 0.65$; full list in Supplementary Table). **c, d,** *t*-SNE plot (**c**) coloured by disease ($n = 6$ non-polyp samples, $n = 6$ polyp samples) and violin plots (**d**) (all violins generated using standard Seurat implementation with default smoothing, density generated at $>25\%$ positive values, widest aspect centre of positive measures, minima and maxima within scale representing all points) for differentially expressed genes across disease state in differentiating or secretory cells. 2,566 cells, $n = 6$ non-polyp samples; 1,796 cells, $n = 6$ polyp samples. Bimodal test, all $P < 2.03 \times 10^{-55}$ or less with Bonferroni correction (exact values in Supplementary Table 3). **e,** *t*-SNE plot of 18,325 re-clustered single cells from merged nasal scrapings ($n = 9$) and surgical samples ($n = 12$) by location (left) (healthy InfTurb (3,681 cells, $n = 3$ samples), polyp-bearing patient InfTurb (1,370 cells, $n = 4$ samples), non-polyp EthSin surgical samples (5,928 cells, $n = 6$ samples), and polyp surgical and scraping samples directly from polyp in EthSin (7,346 cells, $n = 8$ samples)), and cell type (right; for immune cells, see Extended Data Fig. 6) (3,152 basal, 3,089 differentiating, 8,840 secretory, 1,105 ciliated, and 2,139 glandular cells). **f,** Heat map of secretory cells (1,000 cells per location) displaying select genes ($AUC > 0.65$; Supplementary Table 3). **g,** Violin plots of IFN- α , IFN- γ , and IL-4/IL-13 gene signatures for secretory cells. Healthy InfTurb (3,414 cells), polyp-bearing patient InfTurb (1,239 cells), non-polyp EthSin surgical samples (1,048 cells), polyp surgical and scraping samples directly from polyp in EthSin (3,139 cells); effect size -1.16 (IFN- α), -1.05 (IFN- γ) and 1.32 (IL-4/IL-13), polyp EthSin versus healthy. Mann-Whitney *U*-test, $P < 2.2 \times 10^{-16}$.

cells to identify where basal cells become ‘stuck’ in polyps. In the non-polyp ecosystem, we observed that basal cells traverse a wider swath of common pseudotime, with the majority of secretory cells distributed towards the trajectory’s terminus (Fig. 4d, e, Extended Data Fig. 9a). Conversely, in polyps, basal cells accumulate shy of the trajectory’s midpoint, losing the true progenitor position occupied by cluster 8, yet failing to contribute towards later differentiation states (Fig. 4d, e, Extended Data Fig. 9a). Ordering cells along this common axis, we identified several genes that were dysregulated in polyps during epithelial cell differentiation (Extended Data Fig. 9b, Supplementary Table 3; Supplementary Discussion XVI).

As these data highlighted an impairment in differentiation of basal cells in polyp tissue, we sorted basal cells (Extended Data Fig. 7h) from three non-polyp and seven polyp tissues and performed Omni assay for transposase-accessible chromatin (ATAC) sequencing (Omni-ATAC-seq) to identify intrinsic epigenetic changes from the integration of extrinsic cellular signalling events²⁸, and subsequent bulk RNA-seq to confirm and extend our findings (Methods). Polyp basal cells were enriched in peaks for bZIP transcription factor motifs, including various AP-1 family members¹¹ such as JUN, along with FOXA1, ATF3, KLF5 and p63, which have been associated with the maintenance of an undifferentiated state, chromatin opening and oncogenesis (Fig. 4f, Extended Data Fig. 9b–f, Supplementary Table 5; Supplementary Discussion XVII). Clustering of enriched motifs revealed changes in correlation according to disease state (Extended Data Fig. 9c–f; Supplementary Discussion XVII). We further identified expressed candidate transcription factors that may bind to these accessible sites (Fig. 4f, g, Extended Data Fig. 9e, f; Supplementary Discussion XVIII). Collectively, our transcriptomic, pseudotemporal and epigenetic studies led us to hypothesize that during chronic T2I, basal cell differentiation is intrinsically impaired through the influence of extrinsic cues (notably, activation of IL-4/IL-13 and Wnt pathways).

To functionally test for intrinsically altered differentiation potential in vitro, we first seeded basal cells from non-polyp or polyp tissue into air–liquid interface (ALI) cultures (Fig. 5a, Extended Data Fig. 9g, Supplementary Table 3; Supplementary Discussion XIX). Our results suggest that basal cells from polyps can be released from their ‘stuck’ state and differentiate towards a mixed-tissue secretory cell phenotype

non-polyp and polyp basal progenitors^{2,10}, identifying increased expression of transcripts involved in extracellular matrix remodelling and chemo-attraction of effector cells, and a decrease in protease-inhibitor expression and metabolic genes in polyps (Fig. 4a). As some of these upregulated genes are canonical IL-4/IL-13 responsive transcripts⁷, we assessed cytokine-induced gene sets. A combined IL-4/IL-13 signature is strongly induced not only in differentiated polyp epithelium, but also in basal cells, with a large effect size between disease states (Fig. 4b, c, Extended Data Fig. 8a, Supplementary Table 4). IFN- α and IFN- γ signatures—indicative of a type 1 immune module⁶—had small effect sizes (Extended Data Fig. 8b, Supplementary Table 4). Furthermore, from specific hallmark genes, we observed altered balance between Wnt (*CD44*) and Notch (*HEY1*) signalling in polyp epithelium favouring Wnt^{26,27} (Fig. 4a, c, Supplementary Table 4). We further contextualized our basal cell findings by defining alterations in the fibroblast niche that correlate with basal hyperplasia, and identifying significant changes in myeloid and endothelial cell gene expression (Extended Data Figs. 7b, 8c–f; Supplementary Discussion XIV, XV).

Next, we used diffusion pseudotime mapping (see Methods), aligning and reconstructing how basal cells differentiate to mature secretory

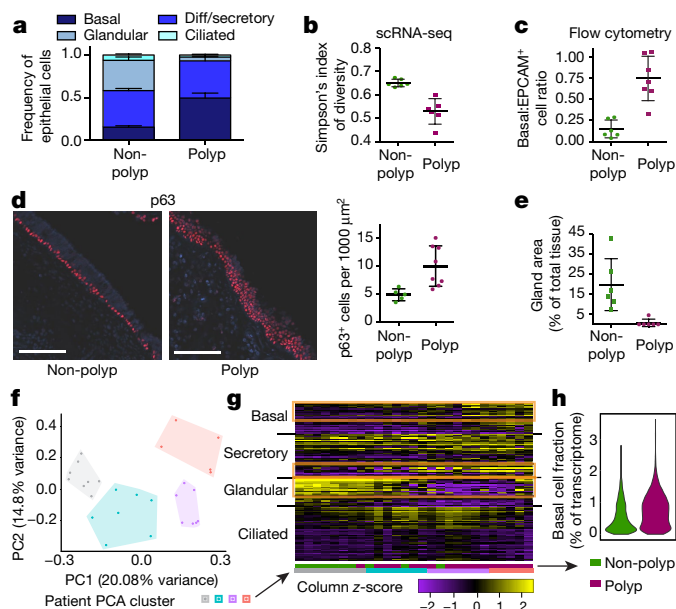


Fig. 3 | Reduced epithelial ecological diversity and basal cell hyperplasia in nasal polyps. **a**, scRNA-seq cell frequency (data from Fig. 2a; Extended Data Fig. 7a shows individual points) calculated for each sample. Basal cell, $P = 0.00023$; glandular, $P < 0.0001$; ciliated, $P = 0.0387$. non-polyp versus polyp. **b**, Simpson's index (Methods, $P = 0.0021$). $n = 6$ non-polyp samples, $n = 6$ polyp samples. Two-sided t -test; data are mean \pm s.e.m. **c**, Quantification of cells by flow cytometry (Extended Data Fig. 7, full gating). $n = 6$ non-polyp samples, $n = 7$ polyp samples. Two-sided t -test; $P = 0.0005$; data are mean \pm s.e.m. **d**, Immunofluorescence (left) and quantification (right) of p63⁺ basal cells normalized to 1,000 μm^2 of epithelium. $n = 5$ non-polyp patients, 10 sections; $n = 8$ polyp patients, 41 sections. Mann–Whitney U -test, $P = 0.0282$. Data are mean \pm s.d. Scale bar, 100 μm . **e**, Quantification of gland area. $n = 6$ non-polyp, $n = 6$ polyp patients. Mann–Whitney U -test, $P = 0.0022$ (Extended Data Fig. 7, isotype and representative). **f**, **g**, Bulk-tissue RNA-seq deconvolution by PCA (**f**) and heat map (**g**) over epithelial subset-specific genes (rows) with k -nearest neighbour (k NN) clusters ($n = 4$; Methods), from $n = 10$ non-polyp samples, $n = 17$ polyp samples (columns). **h**, Violin plot of basal cell gene fraction in scRNA-seq epithelium (Methods, Supplementary Table 4). 5,928 cells, $n = 6$ non-polyp; 4,346 cells, $n = 6$ polyp. Effect size 0.457, Mann–Whitney U -test, $P < 2.2 \times 10^{-16}$.

if provided with strong and sustained extrinsic cues, even in the presence of IL-13 (Fig. 5b; Extended Data Figs. 7h, 9f, h, i; Supplementary Discussion XIX).

Second, since ALI cultures enforced strong terminal differentiation, we directly tested how IL-4/IL-13 act to induce rapid expression of genes in basal cells cultured for 5 weeks ex vivo, hypothesizing that polyp basal cells would respond more vigorously to exogenous cytokines than non-polyp basal cells¹⁴. Surprisingly, we identified 482 genes that were induced in non-polyp basal cells, but only 42 in polyps (Fig. 5c, Supplementary Table 3). Principal component analysis (PCA) highlighted that whereas unstimulated non-polyp basal cells grouped together, polyp basal cells were distributed along PC1, which captured cytokine stimulation (Fig. 5c). Identifying overlaps in genes that are significantly induced by cytokine treatment in non-polyp basal cells with genes that are upregulated at baseline in polyp versus non-polyp basal cells resolved 132 genes (Fig. 5c, Supplementary Table 3).

We focused on the central overlap of these three differential expression tests, which included *CTNNB1* (β -catenin), the key effector of Wnt pathway activation^{26,27} (Fig. 5c). We highlight the fundamental finding that *CTNNB1* was robustly induced in non-polyp and polyp basal cells in a dose-sensitive fashion to IL-4 and IL-13. Moreover, baseline *CTNNB1* expression in polyp basal cells was equivalent to the levels induced by cytokine treatment of non-polyp cells (Fig. 5d). Wnt-pathway target genes were significantly upregulated across the tested doses, confirming overall activation of the pathway, and of specific

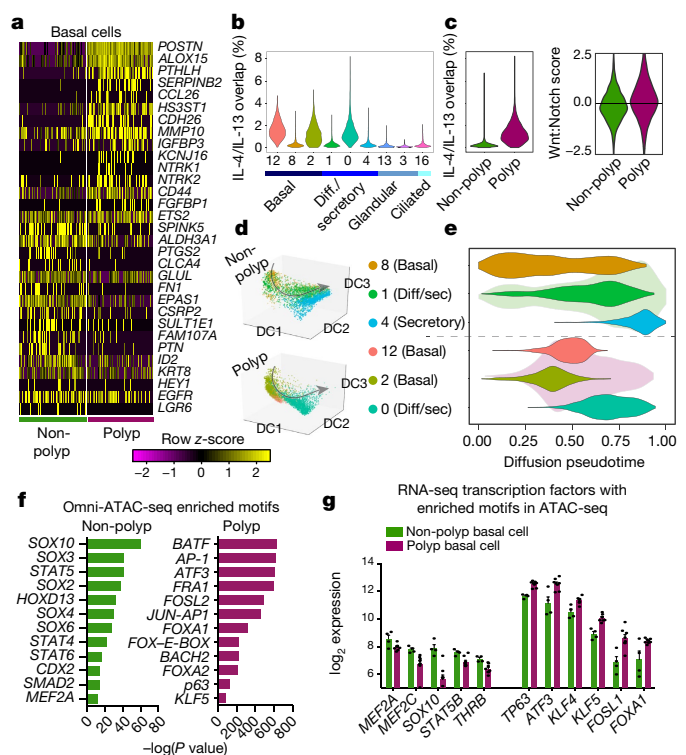


Fig. 4 | T2I cytokines and developmental pathways converge at the epigenetic level in basal cells to intrinsically impair differentiation in vivo. **a**, Heat map of select genes over basal cell clusters 8 and 12. $n = 6$ non-polyp samples, 860 cells; $n = 6$ polyp samples, 858 cells. Bimodal test, all displayed genes $P < 1.97 \times 10^{-39}$ or less with Bonferroni correction (Supplementary Table 3). **b**, Violin plot (Extended Data Fig. 8; cell numbers in Supplementary Table 3, genes listed in Supplementary Table 4) for commonly-induced IL-4/IL-13 gene signature. Mann–Whitney U -test, $P < 1.76 \times 10^{-15}$, relative to mean score, with Bonferroni correction. **c**, Violin plots of shared IL-4/IL-13 signature (Mann–Whitney U -test, $P < 2.2 \times 10^{-16}$, effect size 1.305) and Wnt:Notch target gene ratio. Two-sided t -test, $P < 2.2 \times 10^{-16}$, effect size 0.334. Note the truncated axis; zero indicates equal scores. $n = 6$ non-polyp samples, 5,928 cells; $n = 6$ polyp samples, 4,346 cells. **d**, **e**, Diffusion pseudotime over epithelial cells with unified gene list (**d**) (Supplementary Table 3 cell numbers, gene list) and violin plot of pseudotime component with green ($n = 6$ non-polyp samples) and purple ($n = 6$ polyp samples) underlying distribution (**e**). DC, diffusion component. **f**, **g**, Omni-ATAC-seq and HOMER motif enrichment (**f**) over background peaks. All Q values < 0.0002 , Benjamini correction. Transcription factors from low-input RNA-seq (**g**) on sorted basal cell populations (\log_2 expression log normalized using DESeq2). Two-sided t -test, $P < 0.05$ or less; Holm–Sidak correction; data are mean \pm s.e.m. $n = 3$ non-polyp samples (4 RNA-seq analyses); $n = 7$ polyp samples.

factors (such as *CTGF*) (Fig. 5d, Extended Data Fig. 9j; Supplementary Table 4; Supplementary Discussion XX). On the basis of polyp epithelial gene signatures (Fig. 4c) and our functional testing for IL-4/IL-13 induced genes ‘remembered’ by polyp basal cells (Fig. 5c, d), we propose that chronic IL-4/IL-13 exposure in vivo can lead to persistent expression of Wnt/ β -catenin target genes in a cell-intrinsic fashion, even in the absence of exogenous cytokines.

One polyp patient sampled through scraping commenced treatment with a monoclonal antibody targeting the shared IL-4R α subunit of the IL-4 and IL-13 receptors to treat atopic dermatitis, which provided an opportunity to examine the in vivo relevance of our observational, mechanistic and functional data on how T2I cytokines influence basal cell states (Fig. 5e, Extended Data Fig. 10a, b). We compared cells recovered from pre- and 6-week-post-antibody treatment scrapings, and through surgical intervention at 7 weeks after antibody treatment (Fig. 5e, Extended Data Fig. 10a, b, Supplementary Table 7; Supplementary Discussion XXI). We identified basal cells and

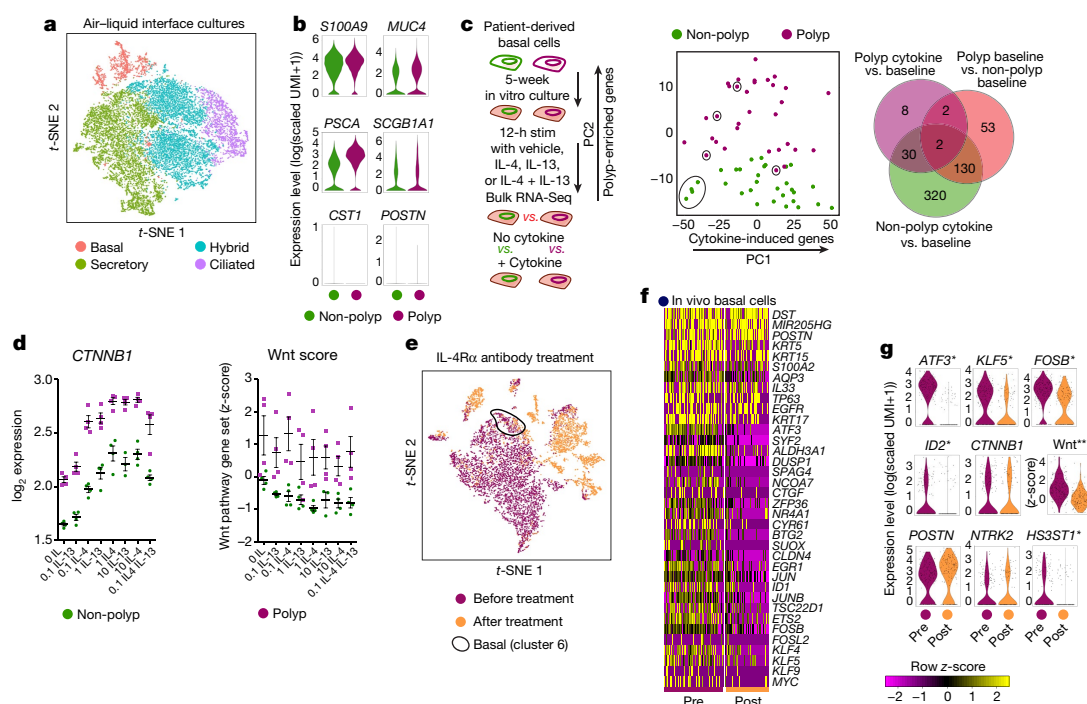


Fig. 5 | Transcriptional memory of IL-4/IL-13 exposure retained by basal cells ex vivo; in vivo IL-4R α blockade partially resets state in polyps. **a**, *t*-SNE plot of cell types from air-liquid interface (ALI) cultures (Extended Data Fig. 9g) over 16,173 single cells ($n = 2$ non-polyp donors, 8,483 cells; $n = 2$ polyp donors, 7,690 cells). **b**, Violin plots for secretory genes (from Fig. 2f) in secretory cells from ALI culture (3,277 non-polyp cells, 3,143 polyp cells). None were statistically significant, except *MUC4*, *PSCA* and *SCGB1A1*, which were more highly expressed in polyp secretory cells. Bimodal $P < 4.04 \times 10^{-16}$, Bonferroni correction. **c**, Basal cell inflammatory memory of IL-4/IL-13 stimulation (left, Methods), displayed as PCA over variable genes (middle; circle denotes no exogenous cytokine) and Venn diagram showing overlaps of differential expression (right). Two-tailed *t*-test, Bonferroni-corrected $P < 0.05$; full gene lists in

Supplementary Table 3. **d**, *CTNNB1* expression and Wnt pathway z-score (Methods) in basal cells from **c**. $n = 4$ samples per dose. Dose expressed in ng ml^{-1} . Two-way ANOVA; $P < 0.0001$ for *CTNNB1*; $P < 0.0282$ for Wnt pathway. $n = 2$ basal cell donors each for non-polyp and polyp samples (**a–d**). **e–g**, scRNA-seq for an individual treated with IL-4R α antibody. *t*-SNE plot of 8,764 single cells from nasal polyps coloured by pre-treatment (5,731 cells) and post-treatment (3,033 cells) samples (**e**) (see also Extended Data Fig. 10), heat map of select genes ($\text{AUC} > 0.68$ core; $P < 2.46 \times 10^{-5}$ or less with Bonferroni correction) over basal cells (**f**) (200 pre, 151 post), and violin plots for select genes (**g**). $*P < 0.00087$, bimodal with Bonferroni correction; Wnt score pre-treatment versus post-treatment, $**P < 2.2 \times 10^{-16}$, two-tailed *t*-test, effect size = 0.942; otherwise not significant. Full gene list in Supplementary Table 3.

generated a heat map containing their top marker genes, agnostic to treatment, followed by genes that were differentially expressed pre- and post-treatment, leveraging myeloid cells to identify basal-specific changes (Fig. 5f, Extended Data Fig. 10a–c, Supplementary Table 3; Supplementary Discussion XXI).

In the context of our earlier findings, these results enabled us to identify several key gene sets, including a conserved core set of basal cell genes (Extended Data Fig. 10d). Several transcription factors that are upregulated in polyp basal cells identified through Omni-ATAC-seq and RNA-seq (*ATF3*, *KLF5* and *FOSB*) were significantly downregulated after treatment (Fig. 5g). Whereas Wnt pathway target gene expression was globally reduced, *CTNNB1* expression was retained, as was expression of genes that are upregulated both in vitro and in vivo in polyp basal cells, suggesting that some disease-associated genes in this patient, and at this time point, persist (Fig. 5g, Extended Data Fig. 10d; Supplementary Discussion XXI).

Lastly, we sought to understand how changes in the basal epithelium propagated through to secretory cells. Within secretory cells recovered from scrapings of both InfTurb and accessible polyp tissue pre- and post-treatment, our data suggest that even though CRS-EthSin samples have unique secretory cell signatures (Fig. 2f), cytokine blockade leads to expression of genes associated with healthy InfTurb secretory cells, even in polyp tissue (Extended Data Fig. 10e–h, Supplementary Table 8; Supplementary Discussion XXI).

One goal of understanding the cellular and molecular pathways activated in T2I is to provide mechanisms that explain persistent chronic allergic inflammatory disease²⁹. Using scRNA-seq applied to patients across the CRS spectrum, our study provides descriptive, mechanistic and functional insights into an enigmatic basal cell state and

productive differentiation of a barrier tissue. We reveal differences in expression of antimicrobial genes by secretory cells relative to healthy tissue, a loss of glandular cell heterogeneity, and strong induction of a transcriptional program by IL-4/IL-13 at the level of basal progenitor cells¹⁵. Our data may help to explain why nasal polyposis is associated with infections by specific microorganisms⁷, and how a monoclonal antibody that targets the shared IL-4/IL-13 receptor can reduce nasal polyp burden.

Together with recent work in the murine intestinal tract and skin^{11,13,16,17,30}, our results provide evidence in humans for the emerging paradigm of stem cell dysfunction altering the set point of barrier tissues, highlighting substantial overlap among putative driving transcription factors (*ATF3*, *AP-1*, *p63* and *KLF5*)¹³. This demonstrates that the principle of inflammatory memory²⁸ underlying barrier tissue adaptation is a generalizable phenomenon that is observed in distinct anatomical locations, inflammatory modules, and species. We build on these findings by culturing basal cells ex vivo and identifying the indelible mark of IL-4/IL-13 as a baseline induction of the Wnt pathway. We propose that basal cells form ‘memories’ of chronic exposure to an inflammatory T2I environment, shifting the entire cellular ecosystem away from productive differentiation and propagating disease. Future work will seek to determine the relative contributions of memory stored in distinct cellular compartments to develop the most effective mechanisms by which to erase them.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0449-8>.

Received: 5 November 2017; Accepted: 4 July 2018;
Published online 22 August 2018.

- Schleimer, R. P. & Berdnikovs, S. Etiology of epithelial barrier dysfunction in patients with type 2 inflammatory diseases. *J. Allergy Clin. Immunol.* **139**, 1752–1761 (2017).
- Hogan, B. L. et al. Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 (2014).
- Whitsett, J. A. & Alenghat, T. Respiratory epithelial cells orchestrate pulmonary innate immunity. *Nat. Immunol.* **16**, 27–35 (2015).
- Iwasaki, A., Foxman, E. F. & Molony, R. D. Early local immune defences in the respiratory tract. *Nat. Rev. Immunol.* **17**, 7–20 (2017).
- Holtzman, M. J., Byers, D. E., Alexander-Brett, J. & Wang, X. The role of airway epithelial cells and innate immune cells in chronic respiratory disease. *Nat. Rev. Immunol.* **14**, 686–698 (2014).
- Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat. Immunol.* **16**, 343–353 (2015).
- Schleimer, R. P. Immunopathogenesis of chronic rhinosinusitis and nasal polyposis. *Annu. Rev. Pathol.* **12**, 331–357 (2017).
- Zhao, L. et al. Increase of poorly proliferated p63⁺/Ki67⁺ basal cells forming multiple layers in the aberrant remodeled epithelium in nasal polyps. *Allergy* **72**, 975–984 (2017).
- Hansel, F. K. Clinical and histopathological studies of the nose and sinuses in allergy. *J. Allergy* **1**, 43–70 (1929).
- Rock, J. R. et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl Acad. Sci. USA* **106**, 12771–12775 (2009).
- Karin, M. & Clevers, H. Reparative inflammation takes charge of tissue regeneration. *Nature* **529**, 307–315 (2016).
- Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
- Naik, S. et al. Inflammatory memory sensitizes skin epithelial stem cells to tissue damage. *Nature* **550**, 475–480 (2017).
- Netea, M. G. et al. Trained immunity: a program of innate immune memory in health and disease. *Science* **352**, aaf1098 (2016).
- Rochman, M. et al. Neurotrophic tyrosine kinase receptor 1 is a direct transcriptional and epigenetic target of IL-13 involved in allergic inflammation. *Mucosal Immunol.* **8**, 785–798 (2015).
- von Moltke, J., Ji, M., Liang, H. E. & Locksley, R. M. TGF- β -cell-derived IL-25 regulates an intestinal ILC2-epithelial response circuit. *Nature* **529**, 221–225 (2016).
- Lindemans, C. A. et al. Interleukin-22 promotes intestinal-stem-cell-mediated epithelial regeneration. *Nature* **528**, 560–564 (2015).
- Cheng, L. E. & Locksley, R. M. Allergic inflammation—innately homeostatic. *Cold Spring Harb. Perspect. Biol.* **7**, a016352 (2014).
- Palm, N. W., Rosenstein, R. K. & Medzhitov, R. Allergic host defences. *Nature* **484**, 465–472 (2012).
- Gieseck, R. L. III, Wilson, M. S. & Wynn, T. A. Type 2 immunity in tissue repair and fibrosis. *Nat. Rev. Immunol.* **18**, 62–76 (2017).
- von Andrian, U. H. & Mackay, C. R. T-cell function and migration. *Two sides of the same coin. N. Engl. J. Med.* **343**, 1020–1034 (2000).
- Allakhverdi, Z. et al. Thymic stromal lymphopoietin is released by human epithelial cells in response to microbes, trauma, or inflammation and potentially activates mast cells. *J. Exp. Med.* **204**, 253–258 (2007).
- Wambre, E. et al. A phenotypically and functionally distinct human T_H2 cell subpopulation is associated with allergic disorders. *Sci. Transl. Med.* **9** (2017).
- Portelli, M. A., Hodge, E. & Sayers, I. Genetic risk factors for the development of allergic disease identified by genome-wide association. *Clin. Exp. Allergy* **45**, 21–31 (2015).
- Zuo, W. L. et al. Ontogeny and biology of human small airway epithelial club cells. *Am. J. Respir. Crit. Care Med.* (2018).
- Boscke, R. et al. Wnt signaling in chronic rhinosinusitis with nasal polyps. *Am. J. Respir. Cell Mol. Biol.* **56**, 575–584 (2017).
- Nusse, R. & Clevers, H. Wnt/ β -catenin signaling, disease, and emerging therapeutic modalities. *Cell* **169**, 985–999 (2017).
- Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
- Lambrecht, B. N. & Hammad, H. The immunology of the allergy epidemic and the hygiene hypothesis. *Nat. Immunol.* **18**, 1076–1083 (2017).
- Beyaz, S. et al. High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. *Nature* **531**, 53–58 (2016).

Acknowledgements We thank S.L. Carroll for technical support with Seq-Well experiments; H. Raff for RNA extraction; J. Lai for histology; A. Chicoine of the Brigham and Women's Human Immunology Flow Core for assistance with isolating cells; L. Ludwig, J. Hammelman and J. Buenrostro for advice on reagents and analysis for ATAC-seq; D. Lingwood, U.H. von Andrian, B. Walker, S. Pillai, N. Yosef, S. Rakoff-Nahoum, S. Beyaz, C. Borges, M.B. Cole, N. Yosef, R. Satija and C. Bingle for discussions and comments on the manuscript; Shalek Laboratory members for experimental and computational advice; and M. Morrison for administrative support. A.K.S. was supported by the Searle Scholars Program, the Beckman Young Investigator Program, the Pew-Stewart Scholars, a Sloan Fellowship in Chemistry, NIH grants 1DP2OD020839, 2U19AI089992, 1U54CA217377, P01AI039671, 5U24AI118672, 2RM1HG006193, 1R33CA202820, 2R01HL095791, 1R01AI138546, 1R01HL126554, 1R01DA046277, 2R01HL095791, and Bill and Melinda Gates Foundation grants OPP1139972 and BMGF OPP1116944; N.A.B. by NIH R01HL120952 and Steven and Judy Kaye Young Innovators Award; T.M.L. by NIH R01HL128241; J.A.B. by NIH U19AI095219, R01AI078908, R01AI136041, R01HL136209; D.F.D. by T32AI007306 (to J.A.B.); K.M.B. by NIH AADCRC Opportunity Fund Award U19AI070535; and K.N.C. by NIH K23AI118804. S.K.N. was supported by NIH 2R01GM081871-09 to B.B. Support was also provided from the Koch Institute Support (core) Grant P30-CA14051 from the NCI, and Ragon Institute NIH-funded Centers for AIDS Research (P30 AI060354, Harvard University Center for AIDS Research), supported by NIH co-funding and participating Institutes and Centers: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH, NIA, FIC, and OAR. J.O.M. is a HHMI Damon Runyon Cancer Research Foundation Fellow (DRG-2274-16), and thanks S. Montanes-Ordovas for encouraging him to work on human allergic disease.

Reviewer information *Nature* thanks R. Schleimer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.O.-M., D.F.D., T.M.L., J.A.B., N.A.B. and A.K.S. designed the study. N.B. performed surgeries. J.O.-M., D.F.D., C.D., M.V., K.M.B., K.N.C. and E.Y. collected patient samples and performed single-cell experiments. J.O.-M., M.V., D.F.D. and E.Y. performed in vitro experiments. M.H.W. and T.K.H. provided the Seq-Well platform and expertise. H.R.K. and E.Y. performed histologic analyses. B.B. provided supervision and analysed epigenetic experiments. J.O.-M., D.F.D., S.K.N. and S.W.K. analysed data. J.O.-M., D.F.D., S.K.N., N.A.B. and A.K.S. interpreted data. J.O.-M., D.F.D., N.A.B. and A.K.S. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0449-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0449-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to N.A.B. or A.K.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Study participants and design for single-cell study from ethmoid sinus tissue.

Subjects between the ages of 18 and 75 years were recruited from the Brigham and Women's Hospital (Boston, Massachusetts) Allergy and Immunology clinic and Otolaryngology clinic between May 2014 and March 2018 (Supplementary Table 1). The Institutional Review Board approved the study, and all subjects provided written informed consent. Ethmoid sinus (EthSin) tissue was collected at the time of elective endoscopic sinus surgery from patients with physician-diagnosed CRS with and without nasal polyps on the basis of established guidelines³¹. Patients with polyps included patients with aspirin-tolerant chronic rhinosinusitis with nasal polyps (CRS polyp) and individuals with aspirin-exacerbated respiratory disease (AERD), both of which are referred to as CRS-EthSin-polyp for the purposes of this study. Patients were suspected of having AERD if they had asthma, nasal polyposis, and a history of respiratory reaction on ingestion of a COX 1 inhibitor, with confirmation via a graded oral challenge to aspirin. Subjects with cystic fibrosis and unilateral polyps were excluded from the study. No distinctions were made between these two disease endotypes in our study, as both present with polyposis, but we present the clinical diagnoses in Supplementary Table 1.

A tissue segment (one per patient) for bulk tissue RNA-seq was immediately placed in RNAlater (Qiagen) for RNA extraction. For patient samples loaded on Seq-Well and for flow-sorting to Omni-ATAC-seq/RNA-seq, tissue was received in-hand, placed in RPMI (Corning) with 10% FBS (ThermoFisher 10082-147) and immediately put on ice for transport. Details of the subjects' characteristics included in scRNA-seq cohort, tissue RNA-seq cohort, and basal cell flow cytometry/ATAC-seq/RNA-seq cohort (including age, gender, medication use, and disease severity) are included in Supplementary Table 1.

Originally, we enrolled a healthy control subject with no known history of CRS or nasal polyposis who was undergoing sinus surgery for concha bullosa. However, this subject upon pathology evaluation was found to have mild eosinophilia. A chart review revealed a history of allergic rhinitis and asthma, and their diagnosis was updated to CRS non-polyp clinically by the surgeon upon follow-up visits so we updated their status accordingly in our study. Additionally, non-polyp patient 6 was sampled twice (denoted as 6A and 6B), representing distinct cells that were captured on two different Seq-Well arrays. As such, they should not be viewed as a technical replicate and are referred to as distinct samples.

Collection of inferior turbinate and nasal polyp samples through nasal scraping.

Nasal samples were collected from the inferior turbinate (InfTurb) of healthy control subjects and from the inferior turbinate and accessible polyp tissue in subjects with CRS-EthSin-polyps using the Rhino-Pro Curette, a sterile, disposable, mucosal collection device, as described^{32,33}. One sample was taken from the right and left mid-inferior portion of the inferior turbinate using a gentle scraping motion. In two subjects with CRS polyp, with accessible nasal polyp tissue, the polyp tissue was sampled using the Rhino-Pro Curette under direct visualization. The nasal scrapings were placed directly in RPMI with 10% FBS and immediately put on ice for transport before loading on Seq-Well arrays. Details of the subjects' characteristics (including age, gender, medication use and disease severity) are included in Supplementary Table 1.

Nasal scraping allows for access to the superficial epithelial cell layer of the inferior turbinate³⁴; by contrast, the surgical resections from EthSin that we utilize as the central dataset of this paper contain both epithelial cells and underlying tissue, including sub-mucosal glands³⁴ (Extended Data Fig. 6c). Since scraping samples a proximal but distinct anatomical location with a distinct technique, in addition to collecting InfTurb scrapings from healthy controls ($n=3$), we also collected InfTurb scrapings from individuals with polyps ($n=4$), and, from two of these individuals, from accessible polyps protruding beyond the middle meatus ($n=2$).

One subject with CRS polyps and co-morbid severe atopic dermatitis was started on dupilumab³⁵, a human monoclonal antibody that binds to the IL-4R α subunit, which is approved for severe atopic dermatitis³⁶, and in a randomized, double-blind, placebo-controlled parallel-group study was shown to significantly reduce endoscopic nasal polyp burden after 16 weeks³⁷. The inferior turbinate and nasal polyp tissue was sampled with the Rhino-Pro Curette pre- and post-treatment with 3 doses of dupilumab, and through endoscopic sinus surgery as noted above.

Tissue digestion. Single-cell suspensions from collected surgical specimens were obtained using a modified version of a previously published protocol³⁸, described below in detail. Each specimen was received directly in hand and processed directly with an average time from patient to loading onto the Seq-Well platform of 3 total hours, and never exceeding 4 h. Surgical specimens were collected into 30 ml of ice cold RPMI (Corning). Specimens were finely minced between two scalpel blades and incubated for 15 min at 37°C in a rotisserie rack with end-over-end rotation in 25 ml digestion buffer supplemented with 600 U/ml collagenase IV (Worthington) and 20 μ g/ml DNase 1 (Roche) in RPMI with 10% fetal bovine serum. After 15 min, samples were triturated five times using a syringe with a 16G needle and returned to the rotisserie rack for another 15 min. At the conclusion of the second digest period, samples were triturated an additional five times using a syringe with

a 16G needle, at which point the digest process was stopped via the addition of EDTA to 20mM. Nasal scrapings were dissociated with one 15 min dissociation using collagenase and the 16G needle trituration was omitted and instead replaced with P1000 pipette trituration, as cell yields were typically <20,000 total cells. Processing downstream remained identical. Samples were typically fully dissociated at this step and were filtered through a 70- μ m cell strainer and spun down at 500g for 10 min followed by a rinse with ice-cold PBS (ThermoFisher 10010023, Ca/Mg-free) to 30 ml total volume. Red blood cells (RBCs) were lysed using ACK buffer (ThermoFisher A1049201) for 3 min on ice to remove RBCs, even if no RBC contamination was visibly seen in order to maintain consistency across patient groups. Cells were then washed with sterile PBS and spun down at 500g for 5 min, resuspended in complete RPMI medium with 2% FCS (RPMI1640 (ThermoFisher 61870-127), 100 U/ml penicillin (ThermoFisher 15140-122), 100 μ g/ml streptomycin (ThermoFisher 15140-122), 10 mM HEPES (ThermoFisher 15630-080), 2% FCS (ThermoFisher 10082-147), 50 μ g/ml gentamicin (ThermoFisher 15750-060)), and counted to adjust concentration to 100,000 cells per ml for loading onto Seq-Well arrays.

Flow cytometry, cell sorting, and analysis. Single-cell suspensions in FACS Buffer (HBSS (ThermoFisher 14170161, Ca/Mg-free) supplemented with 2% FCS) were pre-incubated with Fc-Block (BD 564220) before staining for surface antigens. The following antibodies were used to identify basal cells via flow cytometry: FITC anti-human THY1 (Biolegend, clone 5E10), Brilliant Violet 421 anti-human CD45 (Biolegend, clone HI30), Brilliant Violet 650 anti-human EPCAM (Biolegend, clone 9C4), APC/Cy7 anti-human ITGA6 (Biolegend, clone GoH3), PE/Cy7 anti-human NGFR (Biolegend, clone ME20.4), APC anti-human PDPN (Biolegend, clone NC-08). Cells were stained for 30 min on ice in FACS buffer and then washed for immediate sorting. Cells were sorted on a BD FACSAria Fusion cell sorter using BD FACSDiva software. Up to 10,000 basal cells were sorted into 100 μ l BAM banker (Wako chemicals) for Omni-ATAC-seq and cooled to -80°C using a Mr. Frosty freezing container (Thermo scientific). Samples were stored at -80°C until thawed for Omni-ATAC-seq. For bulk RNA-seq, 1,000 basal cells were sorted directly into 5 μ l TCL buffer (Qiagen). FlowJo v10 by TreeStar was used to generate plots.

Histologic analyses. Biopsies were fixed in 4% paraformaldehyde, embedded in paraffin, and 6- μ m sections were prepared and stained with haematoxylin and eosin for quantification of glandular areas. Photomicrographs encompassing the entire area of each biopsy were taken. Total and glandular areas were measured with ImageJ software and expressed as glandular area as a percentage of total area. For p63 immunofluorescence, sections were quenched for 10 min in 1 mg/ml sodium borohydride in PBS. For antigen retrieval, slides were placed in a Coplin jar with preheated citrate target retrieval buffer (DAKO) at 95°C and transferred to a steamer for 60 min. Slides were cooled for 20 min at room temperature and then transferred to distilled water followed by PBS. Samples were blocked with serum free protein block (DAKO) containing 5% normal donkey serum for 60 min. Samples were incubated overnight at 4°C with purified anti-TP63 antibody (Biolegend, clone W15093A). After three washes in PBS-T, samples were incubated with 1:500 Alexa Fluor 647-conjugated donkey anti-mouse IgG (Jackson immunoresearch, catalogue# 715-605-150) and 1:10,000 Hoechst nuclear dye. Quantification of p63⁺ cells was performed in a blinded fashion and involved counting of p63⁺ nuclei relative to background staining with an isotype control primary antibody. As the epithelium can vary in length across sections, we normalized our quantification of total positive nuclei per 1,000 μ m² area of epithelium as measured in ImageJ and report the final value as p63⁺ cells per 1,000 μ m² of epithelium.

Single-cell RNA-seq with Seq-Well. Once a single-cell suspension was obtained from freshly resected EthSin tissue, or scrapings from InfTurb, we used the Seq-Well platform for massively parallel scRNA-seq to capture transcriptomes of single cells on barcoded mRNA capture beads. Full methods on implementation of this platform are available¹². In brief, 20,000 cells were loaded onto one array preloaded with barcoded mRNA capture beads (ChemGenes). The loaded arrays containing cells and beads were then sealed using a polycarbonate membrane with a pore size of 0.01 μ m, which allows for exchange of buffers but retains biological molecules confined within each nanowell. Subsequent exchange of buffers allows for cell lysis, transcript hybridization, and bead recovery before performing reverse transcription en masse. Following reverse transcription using Maxima H Minus Reverse Transcriptase (ThermoFisher EP0753) and an Exonuclease I treatment (NewEngland Biolabs M0293L) to remove excess primers, PCR amplification was carried out using KAPA HiFi PCR Mastermix (Kapa Biosystems KK2602) with 2,000 beads per 50 μ l reaction volume. Libraries were then pooled in sets of six (totaling 12,000 beads) and purified using Agencourt AMPure XP beads (Beckman Coulter, A63881) by a 0.6X SPRI followed by a 0.7X SPRI and quantified using Qubit hsDNA Assay (Thermo Fisher Q32854). Quality of WTA product was assessed using the Agilent hsD5000 Screen Tape System (Agilent Genomics) with an expected peak >1,000 bp tailing off to beyond 5000 bp, and a small or

non-existent primer peak, indicating a successful preparation. Libraries were constructed using the Nextera XT DNA tagmentation method (Illumina FC-131-1096) on a total of 600 pg of pooled cDNA library from 12,000 recovered beads using index primers with format as in Gierahn et al.¹². Tagmented and amplified sequences were purified at a 0.6X SPRI ratio yielding library sizes with an average distribution of 650–750 base pairs in length as determined using the Agilent hsD1000 Screen Tape System (Agilent Genomics). Two arrays were sequenced per sequencing run with an Illumina 75 Cycle NextSeq500/550v2 kit (Illumina FC-404-2005) at a final concentration of 2.2–2.8 pM. The read structure was paired end with read 1 starting from a custom read 1 primer¹² containing 20 bases with a 12-bp cell barcode and 8-bp unique molecular identifier (UMI) and read 2 containing 50 bases of transcript information.

Single-cell RNA-seq computational pipelines and analysis. Read alignment was performed as in Macosko et al.³⁹. In brief, for each NextSeq sequencing run, raw sequencing data was converted to demultiplexed FASTQ files using bcl2fastq2 based on Nextera N700 indices corresponding to individual samples/arrays. Reads were then aligned to Hg19 genome using the Galaxy portal maintained by the Broad Institute for Drop-Seq alignment using standard settings. Individual reads were tagged according to the 12-bp barcode sequenced and the 8-bp UMI contained in Read 1 of each fragment. Following alignment, reads were binned onto 12-bp cell barcodes and collapsed by their 8-bp UMI. Digital gene expression matrices (for example, cells-by-genes tables) for each sample were obtained from quality filtered and mapped reads, with an automatically determined threshold for cell count. UMI-collapsed data was used as input into Seurat⁴⁰ (<https://github.com/satijalab/seurat>) for further analysis. Before incorporating a sample into our merged dataset, we individually inspected the cells-by-genes matrix of each as a Seurat object.

For analysis of all sequenced surgical ethmoid sinus resection samples, we merged UMI matrices across all genes detected in any condition and generated a matrix retaining all cells with at least 500 UMI detected (19,196 cells and 31,032 genes). This table was then used to set up the Seurat object in which any cell with at least 300 unique genes was retained and any gene expressed in at least 5 cells was retained (Supplementary Information; an R Script is included from this point to set up the Seurat object and walk reader through dimensionality reduction and basic data visualization). The object was initiated with log-normalization, from a UMI + 1 count matrix, scaling, and centering set to true. The total number of cells passing these filters captured across all patients was 18,624 cells with 22,575 genes, averaging 1,503 cells per sample with a range between 789 cells and 3,109 cells (Extended Data Fig. 1a, b, Supplementary Table 2). Before performing dimensionality reduction, data was subset to include cells with less than 12,000 UMI, and a list of 1,627 most variable genes was generated by including genes with an average normalized and scaled expression value greater than 0.13 and with a dispersion (variance/mean) greater than 0.28. We then performed principal component analysis (PCA) over the list of variable genes. For both clustering and *t*-stochastic neighbour embedding (*t*-SNE), we used the first 12 principal components, as upon visual inspection of genes contained within, each contributed to a non-redundant cell type and this reflected the inflection point of the elbow plot. We used FindClusters within Seurat (which utilizes a shared nearest neighbour (SNN) modularity optimization based clustering algorithm) with a resolution of 1.2 and *t*-SNE set to fast with the Barnes–Hut implementation to identify 21 clusters across the 12 input samples.

For analysis of all sequenced InfTurb scraping samples, the object was initiated with log-normalization, from a UMI + 1 count matrix, scaling, and centering set to true. The total number of cells passing these filters captured across all patients was 18,704 cells with 24,842 genes, averaging 2,078 cells per sample with a range between 65 cells and 5,625 cells (note: The 65-cell sample was a very mucus-laden polyp inferior turbinate sample, perhaps explaining the low cell yield, but clustered well within the three other samples each containing 253, 599, and 1,381 cells). Before performing dimensionality reduction, data was subset to include cells with less than 10,000 UMI, and a list of 1,499 most variable genes was generated by including genes with an average normalized and scaled expression value greater than 0.22 and with a dispersion (variance/mean) greater than 0.26. We then performed PCA over the list of variable genes. For both clustering and *t*-SNE, we used the first 16 principal components, as upon visual inspection of genes contained within, each contributed to a non-redundant cell type and this reflected the inflection point of the elbow plot. We used FindClusters (which utilizes an SNN modularity optimization based clustering algorithm) with a resolution of 1 and *t*-SNE set to fast with the Barnes–Hut implementation to identify 18 clusters across the 9 input samples.

For analysis of all sequenced ALI cultures, the object was initiated with log-normalization, from a UMI + 1 count matrix, scaling, and centering set to true. The total number of cells passing these filters captured across all patients was 16,173 cells with 27,396 genes, averaging 2,448 cells per sample with a range between 1,980 cells and 3,009 cells. Before performing dimensionality reduction,

data was subset to include cells with less than 25,000 UMI, and a list of 1,670 most variable genes was generated by including genes with an average normalized and scaled expression value greater than 0.35 and with a dispersion (variance/mean) greater than 0.35. We then performed PCA over the list of variable genes. For both clustering and *t*-SNE, we used the first 16 principal components, as upon visual inspection of genes contained within, each contributed to a non-redundant cell state and this reflected the inflection point of the elbow plot. We used FindClusters (which utilizes an SNN modularity optimization based clustering algorithm) with a resolution of 0.6 and *t*-SNE set to fast with the Barnes–Hut implementation to identify 11 clusters across the 4 input samples.

Cell type identification and within cell type analysis. To identify genes which defined each cluster, we performed a ROC test implemented in Seurat with a threshold set to an area under the curve of 0.65. Top marker genes with high specificity were used to classify cell clusters into cell types (Fig. 1a–c; Extended Data Fig. 1e) based on existing biological knowledge. Three clusters were considered doublets (588 cells) based on co-expression of markers indicative of distinct cell types at ~1/2 the expression level detected in the parent cell cluster (for example, T cell and myeloid cell) and removed from further analyses yielding a matrix with 18,036 cells used in all subsequent steps. Closely related clusters were merged to cell types based on biological curation and analysis of hierarchical cluster trees yielding ten total cell types (Fig. 1a–c; Extended Data Fig. 1e). We identified a much smaller number of eosinophils than expected in our single-cell data. Specifically, if we do not place bulk tissue immediately into RNA-later within 10 min, we cannot reliably detect eosinophil associated transcripts. However, flow cytometrically we recover from 0.5% to 5% of total cells fitting eosinophil profiles from polyps, and focused single-cell studies on granulocytes at the expense of the full ecosystem are possible and the topic of future work (data not shown). With the gentler tissue dissociation required for scrapings, we recovered a greater frequency of eosinophils from polyps in line with flow data (0.31% to 4.6% of cells; Extended Data Fig. 6d). We also did not find a distinct cluster of ILCs, as they are around 0.01 to 0.1% of CD45 cells across the CRS spectrum, per existing literature⁴¹, and extrapolating to the number of CD45 cells we captured, we would have detected between 0.8 and 8 ILCs. To investigate further granularity present within cell types, such as T cells, myeloid cells, fibroblasts, endothelial cells, and epithelial cells, we subset these cells from the Seurat object and re-ran dimensionality reduction and clustering (Extended Data Figs. 3, 4 and 6). The process used for clustering and subset identification was adapted for each cell type to optimize the parameters of variable genes, principal components, and resolution of clusters desired. Canonical correlation analysis⁴² (CCA) was also performed to validate epithelial cell type classification across disease states (Extended Data Fig. 5; Supplementary Information).

Differential expression and fractional contribution of gene set to transcriptome. To identify differentially expressed genes within cell types across non-polyp and polyp disease states, we used the ‘bimod’ setting in FindMarkers implemented in Seurat based on a likelihood ratio test designed for single-cell differential expression incorporating both a discrete and continuous component⁴³. To determine the expression contribution to a cell’s transcriptome of a particular gene list, we summed the total log-normalized expression values for genes within a ‘list of interest’ and divided by the total amount of log-normalized transcripts detected in that cell, giving the proportion of a cell’s transcriptome dedicated to producing those genes. For comparison of Wnt and Notch signalling, we z-scored the expression contribution metric and subtracted the value of Notch from Wnt yielding a metric centred on zero if both scores are equivalent, or weighted in the positive direction if enriched in Wnt. For reference gene lists used, including basal cell⁴⁴, genes induced by IFN- α , IFN- γ , IL-4, IL-13, IL-4/IL-13⁴⁵; Wnt and Notch please see Supplementary Table 4.

Simpson’s index of diversity, and fibroblast gene correlation with basal cell frequency. To measure the ‘richness’ of the epithelial ecosystem⁴⁶, we employed Simpson’s index of diversity (*D*), which we present as $(1 - D)$, and ranges between 0 and 1, with larger values indicating larger sample diversity⁴⁷. We used Simpson’s index to characterize the composition of epithelial cells across basal, differentiating/secretory, glandular, and ciliated groupings in the non-polyp and polyp ethmoid sinus tissue ecosystems, as this metric accounts for both the number of distinct cell types present (for example, species), and the evenness of the cellular composition across those cell types (for example, relative abundance of species to each other). This measure takes into account the total number of members of a cell type, the number of cell types, and the total number of cells present. We calculate $(1 - D)$ for each sample. To determine genes correlated in specific cell types (for example, fibroblasts) with the frequency of basal cells present in a cellular ecosystem, we correlated the average log-normalized single-cell count data for each gene to the rank of samples determined by increasing frequency of basal cells in each ecosystem (8.2% to 19.1% for non-polyp and 27.9% to 70.1% for polyp samples, Extended Data Fig. 7b).

Tissue and sorted basal cell RNA-seq. Population RNA-seq was performed using a derivative of the Smart-Seq2 protocol for single cells⁴⁸. In brief, tissue was collected

directly into RNeasy (Qiagen) in the surgical suite and stored at -80°C until RNA isolation. RNA was isolated from 30 patients using phenol/chloroform extraction and normalized to 5 ng as the input amount for a 2.2X SPRI ratio cleanup using Agencourt RNAClean XP beads (Beckman Coulter, A63987). RNA-seq was performed on a bulk population of sorted basal cells using Smart-Seq2 chemistry, starting with a 2.2X SPRI ratio cleanup. After oligo-dT priming, Maxima H Minus Reverse Transcriptase (ThermoFisher EP0753) was used to synthesize cDNA with an elongation step at 52°C before PCR amplification (15 cycles for tissue, 18 cycles for sorted basal cells) using KAPA HiFi PCR Mastermix (Kapa Biosystems KK2602). Sequencing libraries were prepared using the Nextera XT DNA tagmentation kit (Illumina FC-131-1096) with 250 pg input for each sample. Libraries were pooled post-Nextera and cleaned using Agencourt AMPure SPRI beads with successive 0.7X and 0.8X ratio SPRI and sequenced with an Illumina 75 Cycle NextSeq500/550v2 kit (Illumina FC-404-2005) with loading density at 2.2 pM, with paired end 35 cycle read structure. Tissue samples were sequenced at an average read depth of 7.98 million reads per sample and 3 samples not meeting quality thresholds were excluded from further analyses yielding 27 total useable samples. Sorted basal cell samples were sequenced at an average read depth of 21.15 million reads per sample and all samples met quality thresholds regarding genomic and transcriptomic alignment.

Tissue and sorted basal cell RNA-seq data analysis. Tissue and sorted basal cell samples were aligned to the Hg19 genome and transcriptome using STAR⁴⁹ and RSEM⁵⁰. Three samples were excluded for low transcriptome alignment ($<25\%$), so we retained 27 samples for further analyses. Differential expression analysis was conducted using DESeq2 package for R⁵¹. Genes regarded as significantly differentially expressed were determined based on an adjusted P value using the Benjamini–Hochberg procedure to correct for multiple comparisons with a false discovery rate <0.05 . We performed ingenuity pathway analysis (IPA, Qiagen) through an instance available through the Broad Institute on the top 1,000 differentially expressed genes (all adjusted $P < 0.05$) from our DESeq2 analysis, taking into account corresponding log-fold change for each gene. We also subset the tissue RNA-seq matrix based on genes found in Supplementary Table 3, which, from our single-cell marker discovery, were specific for basal, differentiating/secretory, glandular, or ciliated cells. We then ran PCA and kNN clustering implemented in R over these genes in order to identify the greatest vectors of variance across samples within the epithelial cell compartment (Fig. 3f, g).

For re-analysis of published data, we used two publicly-available RNA-seq datasets: one profiling normal human olfactory mucosa and the other assessing differences in gene expression between healthy, non-eosinophilic nasal polyps and eosinophilic nasal polyps^{7,52,53}. Note that analysis is done on a per sample basis and as such no comparisons are made across the datasets or samples.

Diffusion pseudotime mapping for differentiation analysis. Using diffusion pseudotime⁵⁴ mapping, which seeks to provide the most likely reconstruction for the developmental progression of a set of cells we built a trajectory for cells within basal and differentiating/secretory epithelial clusters (non-polyp clusters: 8-basal, 1-differentiating/secretory, 4-secretory; and polyp clusters: 12-basal, 2-basal, 0-differentiating/secretory; running several iterations starting from a random seed cell in cluster 8), over the combined basal and apical marker gene list (Fig. 4d; Extended Data Fig. 9a, Supplementary Table 3). By calculating a pseudotime trajectory for cells from both non-polyps and polyps together, we were then able to ask where cells from each disease state fall along a shared inferred temporal axis (Fig. 4d, e; Extended Data Fig. 9a). Diffusion pseudotime⁵⁴ was calculated using the scanpy Python package ‘dpt’ function on log-normalized data for clusters 8, 1 and 4 (predominantly non-polyp, Supplementary Table 3) and 12, 2, and 0 (predominantly polyp, Supplementary Table 3) together. A random root cell was chosen from cluster 8, as this was the basal cell cluster representative of the non-polyp (for example, less aberrant) state, and we also ran iterations with random root cells chosen from the entire set of clusters and it assigned cluster 8 as the cluster most enriched at the beginning of the diffusion map, regardless. Plots were created with the seaborn, matplotlib and pandas packages. Pearson correlations were then calculated for all genes in all cells tested, or for all genes in non-polyp cells and all genes in polyp cells, relative to pseudotime (Extended Data Fig. 9b). Differential correlation testing was performed using the cocor package to identify significance for the difference between correlation coefficients using Fisher’s 1925 z -statistic, accounting for number of cells.

Epigenetic profiling of basal cells using Omni-ATAC-seq. Accessible chromatin profiling⁵⁵ using the Omni-ATAC-seq protocol as described in Corces et al.⁵⁶ was performed on basal cells stored in 100 μl BAMBanker freezing media from 12 patients ($n = 4$ non-polyp (3 retained after data quality filtering) and $n = 8$ polyp). Cells (ranging from 1,000 to 10,000) were thawed quickly in a 37°C rock bath and 900 μl of ice-cold PBS supplemented with Roche Complete Mini Protease inhibitor was added immediately. Cells were split into two 1.5-ml Eppendorf DNA lo-bind tubes to serve as technical replicates. Cells were centrifuged at 500g for 5 min at 4°C , washed once in PBS with protease inhibitor, centrifuged at 500g

for 5 min at 4°C and supernatant was removed completely using two separate pipetting steps with extreme caution taken to avoid resuspension (for example, smooth and consistent aspiration). The transposition reaction consisted of 20- μl total volume of the following mixture (10 μl $2\times$ TD Buffer, 1 or 0.5 μl TDEnzyme, 0.1 μl of 2% digitonin, 0.2 μl of 10% Tween 20, 0.2 μl of 10% NP40, 6.6 μl of $1\times$ PBS and 2.3 μl of nuclease-free water). We performed replicates with two distinct concentrations of TDE since, when dealing with minute clinical samples, flow sorting can sometimes give variable cell numbers, and the ratio of TDE to cells is critical in determining the frequency with which cuts are made in the genome. We optimized in pilot experiments that for basal cell inputs in the range of 500 to 10,000 cells, the aforementioned two ratios gave expected patterns of nucleosome banding in gels (data not shown). We performed two reactions and then later, during in silico analysis, pooled peaks together for downstream analysis. The cells were resuspended into the transposition mixture and incubated at 37°C for 30 min in an Eppendorf Thermomixer with agitation at 300 r.p.m. Transposed DNA was purified using a Qiagen MinElute Reaction Cleanup Kit with elution in 15 μl . Libraries were constructed from 10 μl of DNA using a 50 μl total reaction volume of NEB HF $2\times$ PCR Master Mix with custom Nextera N700 and N500 index primers to barcode samples (also used in Smart-Seq2 protocol). We performed 14 cycles of PCR amplification and SPRI purified at $1.8\times$ ratio. Based on the molarity of each library, we adjusted the number of subsequent PCR cycles to either 3, 4 or 5 more for each sample. We then performed a $0.25\times$ reverse SPRI to remove larger fragments followed by a $1.7\times$ SPRI to purify libraries for sequencing. Libraries were sequenced on an Illumina NextSeq with paired-end 38-cycle read structure at a loading density of 1.95 pM.

Omni-ATAC-seq data analysis. Reads were aligned using bowtie2 using the following flags: ‘-S -p 1 -X 2000 --chunkmbs 1000’ then bam files were created using samtools view with the following flags: ‘samtools view -bS -F 4’. Duplicates were removed with picard. Forward reads were shifted 4 bp and negative reads were shifted 5 bp using a custom Python script and the pysam package as is recommended for ATAC-seq data. Samples for each patient were merged using samtools merge and all patients were downsampled to 3 million reads using custom python scripts and ‘samtools view’ with the ‘-b’ and ‘-s’ flags. MACS2 ‘callpeak’ command was used to call peaks on each sample with flags ‘-f BAMPE -q 0.001 --nomodel --shift -100 --extsize 200 -B --broad’. Peaks from all samples were merged into one peakfile with bedtools and counts of reads per peak for each sample were generated with bedtools multicov. DESeq2 was run with the design ~polyp, testing for significant differences between polyp and non-polyp samples on this peak matrix and differential peaks with Benjamini–Hochberg adjusted P value less than 0.01 with ‘greater’ or ‘less’ null hypotheses were used in downstream analysis. Homer2 was run for known motif finding on differential peaks with the set of all peaks as background⁵⁷. To determine a false discovery rate, Homer2 was run on sets of random peaks chosen with replacement from the set of all peaks.

Epithelial cell culture. Tissues were digested as described above from either non-polyp or polyp surgical resections from the ethmoid sinus. 1,000,000 digested cells were added to a 25- cm^2 tissue culture flask (Corning) pre-coated with 0.03 mg/ml Type I bovine collagen solution (StemCell Technologies) and cultured in PneumaCult-Ex media (StemCell Technologies, 05008). Media was changed every second day until cells reached confluence. Cells were subsequently frozen in 70% basal media with 20% FBS and 10% DMSO.

Air–liquid interface cultures. For air–liquid interface (ALI) cultures⁵⁸, 100,000 cultured epithelial cells per well were added to 0.4- μm pore 24-well polyester membrane inserts (Corning) pre-coated with 0.03 mg/ml Type I bovine collagen solution (StemCell Technologies) with Pneumacult-Ex media (StemCell Technologies, 05008) on both sides of the membrane. After 24 h, apical media was changed to remove dead cells. After 72 h, apical media was removed completely and basal media was changed to Pneumacult-ALI (StemCell Technologies, 05001) supplemented with 5 ml $100\times$ penicillin-streptomycin (Fisher), 1 ml $500\times$ gentamicin/amphotericin B (ThermoFisher), 1 ml 0.2% heparin sodium salt in PBS (StemCell Technologies) and 2.5 ml $200\times$ hydrocortisone stock solution (StemCell Technologies) and 0, 0.1, 1 or 10 ng/ml IL-13 (Biolegend). Basal media was changed every 2–3 days for 21 days, after which membranes were removed and cells dissociated with Stempro Accutase Cell Dissociation Reagent (Gibco) for Seq-Well or flow cytometry. After following scRNA-seq data analysis pipelines described above, cell states recovered in ALI cultures (Fig. 5a; Extended Data Fig. 9g) were related to in vivo cell types⁵⁹.

Basal cell stimulation. Basal cells from non-polyp or polyp surgical resections from ethmoid sinus were placed into epithelial cell culture (for example, ‘lateral expansion’ in the absence of differentiation, see above, inspired by experiments in microglia⁶⁰), passaged, and 10,000 cells seeded at passage 5 (for example, 5 weeks ex vivo) and cultured at confluence in 96-well flat-bottom collagen-coated tissue culture plates (Corning, 3799) for 48 h in Pneumacult-Ex serum-free media (StemCell Technologies, 05008). Cytokines were added for 12 h overnight at increasing doses (0, 0.1, 1, 10 ng/ml) of IL-4 (Biolegend 766205), IL-13 (Biolegend 571104), and

(0.1, ng/ml) IL-4 + IL-13 in combination ($n = 32$ samples non-polyp and $n = 32$ samples polyp basal cells over all conditions, each condition run as a biological duplicate, and a technical duplicate therein), before lysis using RLT + 1% BME (Qiagen and Sigma, respectively). Bulk RNA-seq was performed as described for sorted basal cells starting from lysates. Basal cell stimulation samples were sequenced at an average read depth of 3 million reads per sample and all samples met quality thresholds regarding genomic and transcriptomic alignment.

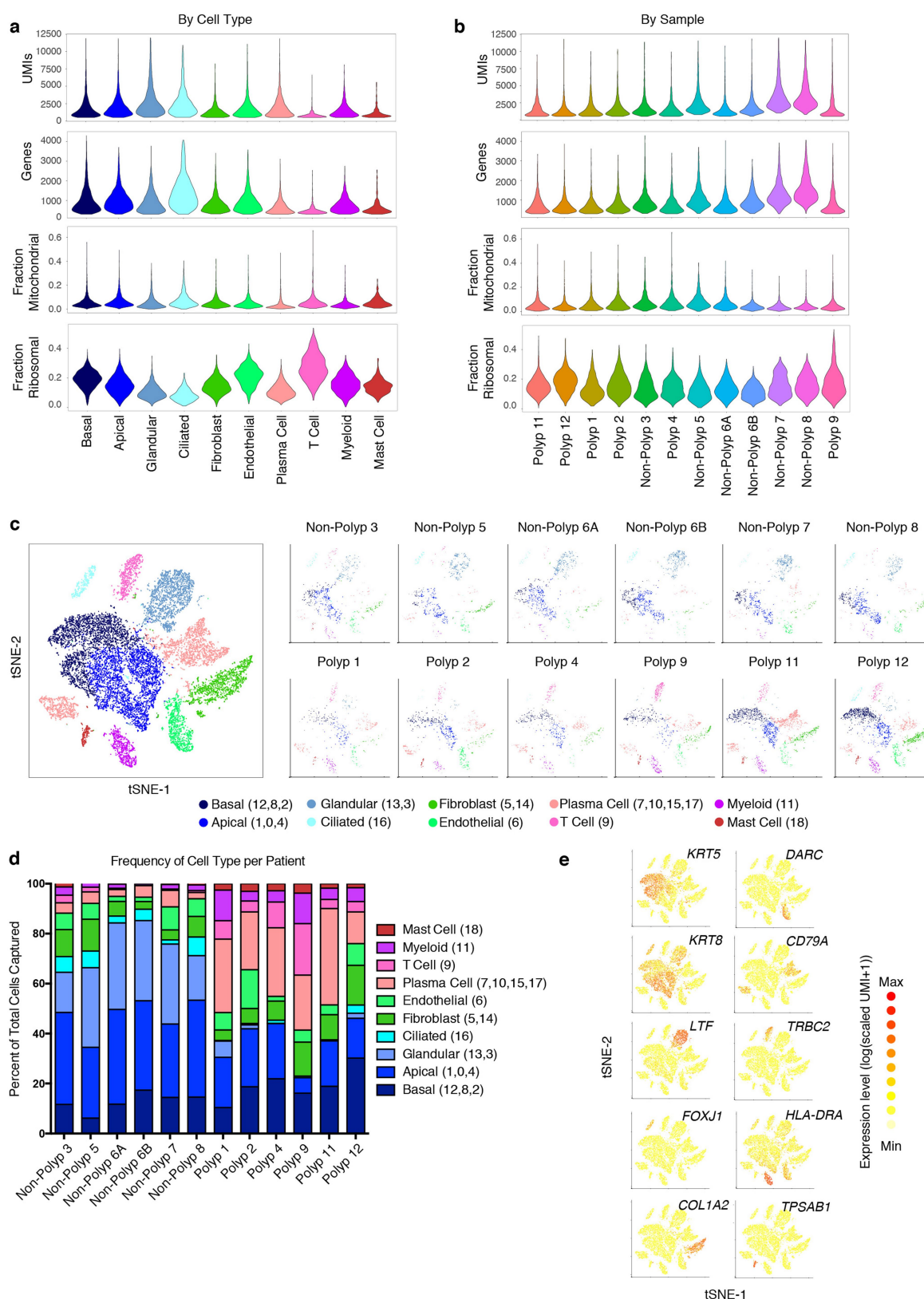
Statistical analyses. No statistical methods were used to predetermine sample size. Number of samples included in analyses are listed throughout figure legends and all represent distinct biological samples. The same surgeon performed surgeries on all individuals and was blinded to study design. The same allergist/immunologist performed nasal scrapings on all samples and was blinded to study design. Quantification of histological sections was performed in a blinded fashion. No samples or cells meeting quality thresholds were excluded from analyses. Where single-cell data was analysed on a gene level, the statistics were performed over the number of cells. Statistical analyses were performed using GraphPad Prism v7.0a, Seurat 1.4.0.1 implemented in RStudio, DESeq2 1.10.1 package implemented in RStudio, and Ingenuity Pathway Analysis run through the Broad Institute, and macs2, DESeq2 and Homer2 for Omni-ATAC-seq. All violin plots were generated using standard Seurat code without modification to smoothing or density. Violin density only generated when >25% of cells in indicated sample have non-zero measurement for gene, widest aspect represents centre of positive measures, minima and maxima are represented within the scale with minima at 0 and maxima encompassing all points for the count-based expression level ($\log(\text{scaled UMI} + 1)$) of each gene. Exact values for all genes displayed and tested available in Supplementary Table 3 organized by panel. All violin plots contain at minimum 100 individual cells in any one cluster (Supplementary Table 3 for precise numbers of cells per cluster and type, most are included in figure legends where space allows), and have points suppressed for ease of legibility. Some violin plots with less than 100 cells have individual data points displayed and corresponding statistical metrics are available in accompanying figure legend and Supplementary Table 3. As some scores followed non-normal distributions as tested for using a Lilliefors normality test, we used a Mann–Whitney U -test where indicated for determining statistical significance. For scores in single-cell data, we report effect sizes in addition to statistical significance as an additional metric for the magnitude of the effect observed. The calculation was performed as Cohen's d where: effect size $d = (\text{Mean}_1 - \text{Mean}_2) / (\text{s.d. pooled})$. Unpaired two-tailed t -tests for direct comparisons and t -test with Holm–Sidak correction, Bonferroni correction, or Benjamini–Hochberg for multiple comparisons, depending on software package used, where appropriate. Mann–Whitney U -test for quantification of histological data due to non-normally distributed data. Pearson correlation thresholds were determined as significant through determination of asymptotic P values through use of `rcorr` function in `Hmisc`, but exact corrected P values by Holm–Sidak method for multiple comparisons are calculated for those highlighted in text using `RcmdrMisc` package. Comparison of Pearson correlation coefficients in pseudotime analyses was done using Fisher's 1925 z -statistic accounting for the number of cells.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The cells-by-genes matrix generated from EthSin surgical resections and analysed during the current study is available along with the manuscript as Supplementary Table 2 alongside R code for standard implementation of Seurat. A cells-by-genes matrix from InfTurb and polyp scraping data are also available as Supplementary Table 6. Dupilumab treatment cells-by-genes matrices are shown in Supplementary Tables 7 and 8. A metadata table encompassing all scRNA-seq samples is provided as Supplementary Table 9. The count and TPM matrices and associated metadata from bulk tissue RNA-seq are available as Supplementary Tables 10, 11, and 12. FASTQ file format data have been deposited in and are available from the dbGaP database under dbGaP accession 30434 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs030434.v1.p1). Marker gene lists for cell types identified in Fig. 1a, b and from resultant analyses in Fig. 2b, for frequencies of cell clusters and types in Fig. 2c, for cell types identified in Fig. 2e, f, 3g, 5a, e, Extended Data Fig. 3a–c, 4c, 5e, 6b, d, 10a, selected comparisons of differential expression in Figs. 2d, 4a, 5c, f, Extended Data Fig. 2c, 10h, and pseudotime correlation in Extended Data Fig. 9b are available as tabs in Supplementary Table 3. Differential peak calling from epigenetic profiling is

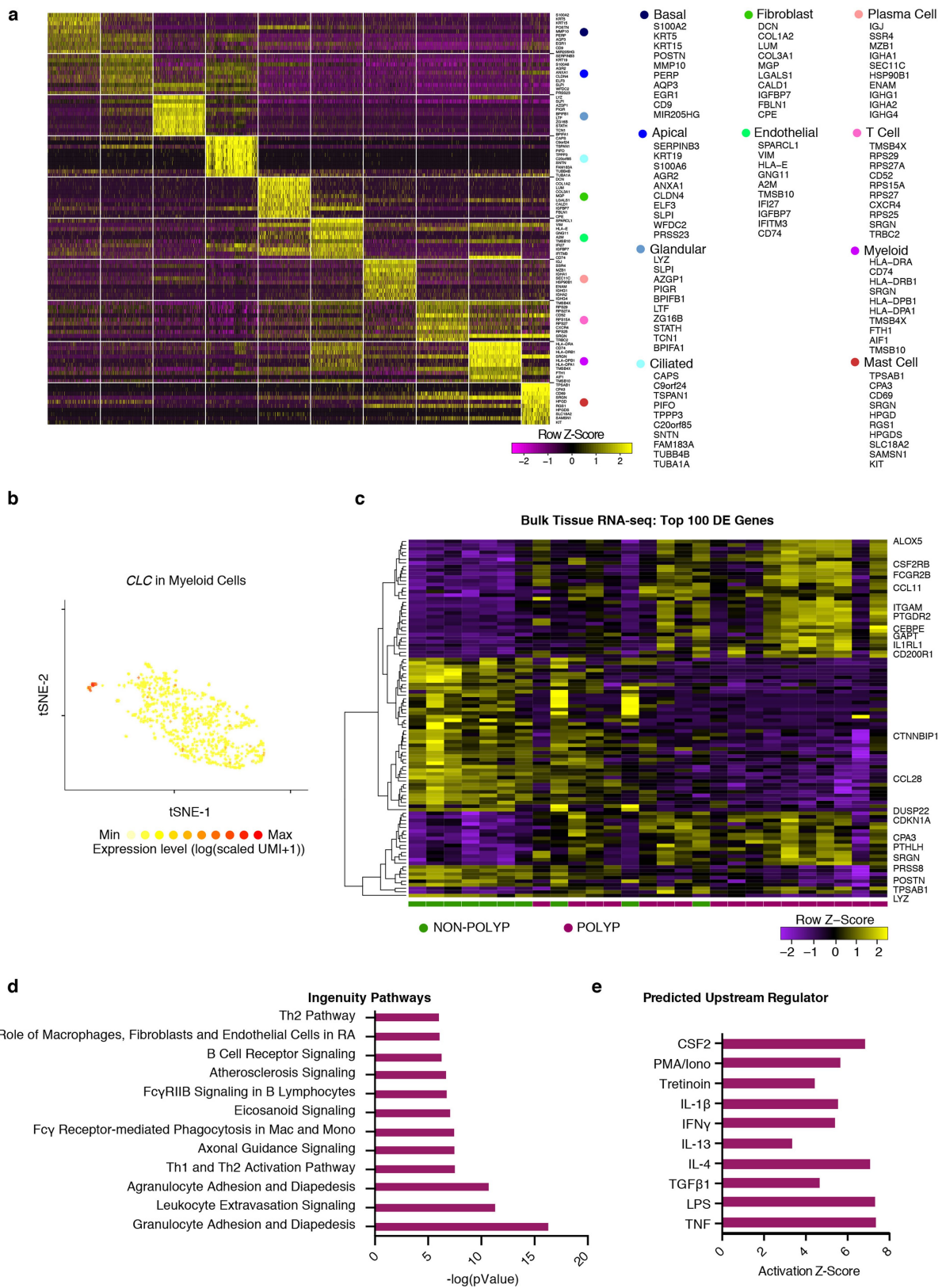
available in Supplementary Table 5. Additional R code for analyses is available at <http://shaleklab.com/resources/>. Aligned and quality-filtered data and complete statistical outputs for the figures are included as Supplementary Tables, with further information at <http://shaleklab.com/resources/>.

31. Meltzer, E. O. et al. Rhinosinusitis: establishing definitions for clinical research and patient care. *J. Allergy Clin. Immunol.* **114**, 155–212 (2004).
32. Dhariwal, J. et al. Mucosal type 2 innate lymphoid cells are a key component of the allergic response to aeroallergens. *Am. J. Respir. Crit. Care Med.* **195**, 1586–1596 (2017).
33. Proud, D., Sanders, S. P. & Wiehler, S. Human rhinovirus infection induces airway epithelial cell production of human β -defensin 2 both in vitro and in vivo. *J. Immunol.* **172**, 4637–4645 (2004).
34. Pipkorn, U. & Karlsson, G. Methods for obtaining specimens from the nasal mucosa for morphological and biochemical analysis. *Eur. Respir. J.* **1**, 856–862 (1988).
35. Wenzel, S. et al. Dupilumab in persistent asthma with elevated eosinophil levels. *N. Engl. J. Med.* **368**, 2455–2466 (2013).
36. Beck, L. A. et al. Dupilumab treatment in adults with moderate-to-severe atopic dermatitis. *N. Engl. J. Med.* **371**, 130–139 (2014).
37. Bachert, C. et al. Effect of subcutaneous dupilumab on nasal polyp burden in patients with chronic sinusitis and nasal polyposis: a randomized clinical trial. *J. Am. Med. Assoc.* **315**, 469–479 (2016).
38. Dwyer, D. F., Barrett, N. A., Austen, K. F. & Immunological Genome Project Consortium. Expression profiling of constitutive mast cells reveals a unique identity within the immune system. *Nat. Immunol.* **17**, 878–887 (2016).
39. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
40. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
41. Poposki, J. A. et al. Group 2 innate lymphoid cells are elevated and activated in chronic rhinosinusitis with nasal polyps. *Immun. Inflamm. Dis.* **5**, 233–243 (2017).
42. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
43. McDavid, A. et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).
44. Hackett, N. R. et al. The human airway epithelial basal cell transcriptome. *PLoS One* **6**, e18378 (2011).
45. Giovannini-Chami, L. et al. Distinct epithelial gene expression phenotypes in childhood respiratory allergy. *Eur. Respir. J.* **39**, 1197–1205 (2012).
46. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368**, 734–737 (1994).
47. Simpson, E. H. Measurement of diversity. *Nature* **163**, 688 (1949).
48. Trombetta, J. J. et al. Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1–4.22.17 (2014).
49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
51. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
52. Olender, T. et al. The human olfactory transcriptome. *BMC Genomics* **17**, 619 (2016).
53. Wang, W. et al. Transcriptome analysis reveals distinct gene expression profiles in eosinophilic and noneosinophilic chronic rhinosinusitis with nasal polyps. *Sci. Rep.* **6**, 26604 (2016).
54. Haghverdi, L., Buttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
55. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
56. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
57. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
58. Fulcher, M. L., Gabriel, S., Burns, K. A., Yankaskas, J. R. & Randell, S. H. Well-differentiated human airway epithelial cell cultures. *Methods Mol. Med.* **107**, 183–206 (2005).
59. Mead, B. E. et al. Harnessing single-cell genomics to improve the physiological fidelity of organoid-derived cell types. *BMC Biol.* **16**, 62 (2018).
60. Gosselin, D. et al. An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, eaal3222 (2017).



Extended Data Fig. 1 | Consistency of cell capture and identification in surgical EthSIN scRNA-seq patient cohort. **a**, Number of unique molecular identifiers (nUMI) and genes identified, and fraction of reads mapping to mitochondrial or ribosomal genes across recovered cell types: 3,222 basal cells, 4,362 apical cells, 2,192 glandular cells, 498 ciliated cells, 835 T cells, 2,976 plasma cells, 1,724 fibroblasts, 1,143 endothelial cells, 811 myeloid cells and 273 mast cells. **b**, nUMI and genes identified, and fraction of reads mapping to mitochondrial or ribosomal genes across patient samples: 789 polyp 1 cells, 1,309 polyp 2 cells, 1,153 polyp 3 cells,

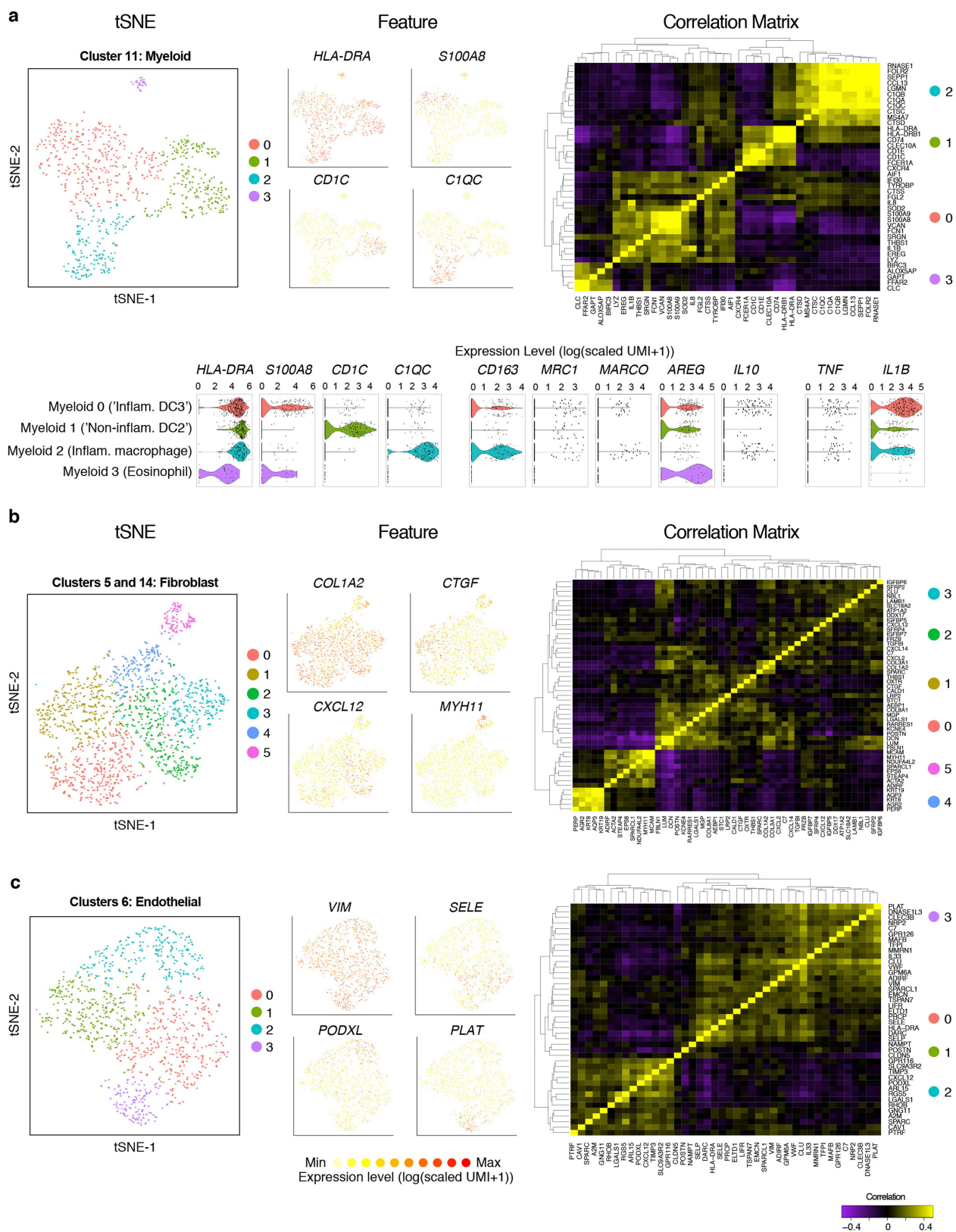
913 polyp 4 cells, 1,219 polyp 5 cells, 1,141 polyp 6A cells, 1,334 polyp 6B cells, 1,314 polyp 7 cells, 1,286 polyp 8 cells, 1,481 polyp 9 cells, 2,988 polyp 11 cells, 3,109 polyp 12 cells. **c**, t-SNE plot as in Fig. 1b coloured by cell types across all patients and then separated by sample: 18,036 single cells ($n = 12$ samples). **d**, The percentage of each cell type recovered within each sample. **e**, Select marker gene overlays displaying binned count-based UMI-collapsed expression level (log(scaled UMI + 1)) on a t-SNE plot from Fig. 1b for key cell types identified (see Supplementary Table 3 for full gene lists); AUC 0.998 to 0.7 for all markers displayed.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Top marker genes for cell types by scRNA-seq and bulk tissue RNA-seq from EthSin recovers expected T2I and eosinophilic modules. **a**, Row-normalized heat map of the top 10 marker genes identified by ROC test ($AUC > 0.73$ for all) over all cell types (Fig. 1b, c) with select genes displayed on y axis and cells on x axis (see Supplementary Table 3 for full gene lists); maximum 500 cells per type. **b**, An overlay of *CLC* (a pathognomonic gene for eosinophils) displaying binned count-based expression level ($\log(\text{scaled UMI} + 1)$) amongst myeloid cells. 811 myeloid cells from $n = 12$ samples. **c**, A row-normalized and row-clustered heat map over the top 100 positively and negatively differentially-expressed genes (50 in each direction) in bulk tissue

RNA-seq of 27 samples from non-polyp ($n = 10$) and polyp ($n = 17$) tissue with select genes displayed. DESeq2 Wald test, all $P < 9.03 \times 10^{-5}$ for genes displayed, corrected for multiple comparisons by Benjamini procedure, samples ordered as in Fig. 3g (see Supplementary Table 4 for full gene list and associated statistics). **d**, The top differentially regulated pathways identified by ingenuity pathway analysis (see Methods) over the top 1,000 differentially expressed genes, as determined by $P < 0.05$ corrected for multiple comparisons by Benjamini procedure, across polyp and non-polyp tissue. **e**, Predicted upstream regulators based on differentially expressed gene modules in polyp tissue relative to non-polyp determined using ingenuity pathway analysis (see Methods).

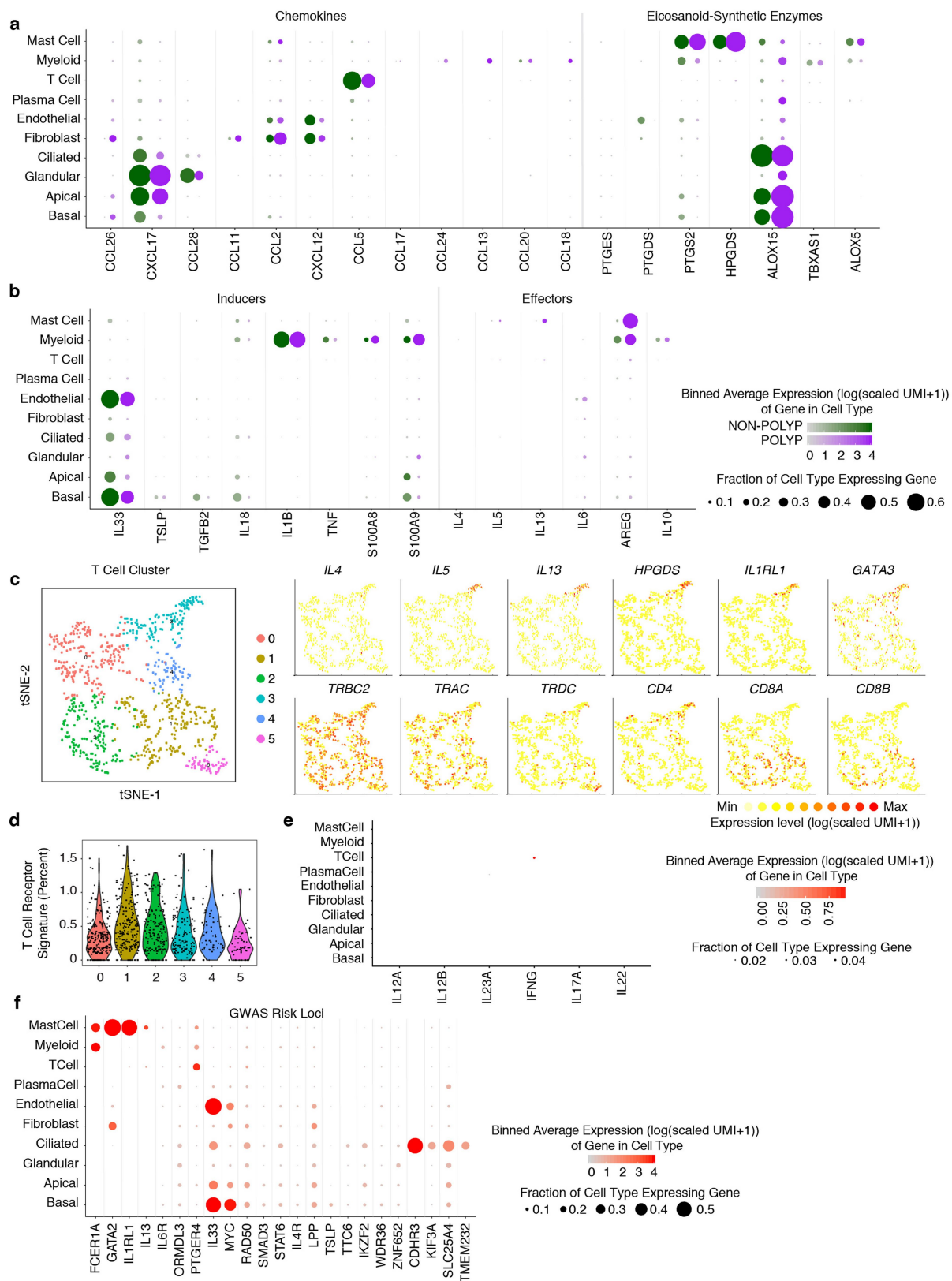


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Sub-clustering of myeloid, fibroblast and endothelial cell types from the EthSin T2I inflammatory ecosystem.

a, *t*-SNE plot of 811 myeloid cells ($n = 6$ non-polyp, $n = 6$ polyp samples), coloured by clusters identified through shared nearest neighbour (SNN) analysis (Supplementary Table 3; Methods), from CRS-EthSin; select marker gene overlays displaying count-based (UMI-collapsed) expression level ($\log(\text{scaled UMI} + 1)$) on a *t*-SNE plot (see Supplementary Table 3 for full gene lists; genes identified by ROC test with AUC 0.689 for *S100A8*, 0.763 for *CD1C*, 0.927 for *CIQC*); a clustered correlation matrix of marker genes identified in single-cell data from myeloid cells; and violin plots for the expression value ($\log(\text{scaled UMI} + 1)$) of selected markers of myeloid activation state. **b**, *t*-SNE plot of 1,724 fibroblasts ($n = 6$ non-polyp, $n = 6$ polyp samples), coloured by clusters identified through shared nearest neighbour (SNN) analysis (Supplementary Table 3; Methods), from CRS-EthSin; select marker gene overlays displaying count-based (UMI-collapsed) expression level ($\log(\text{scaled UMI} + 1)$) on a *t*-SNE plot

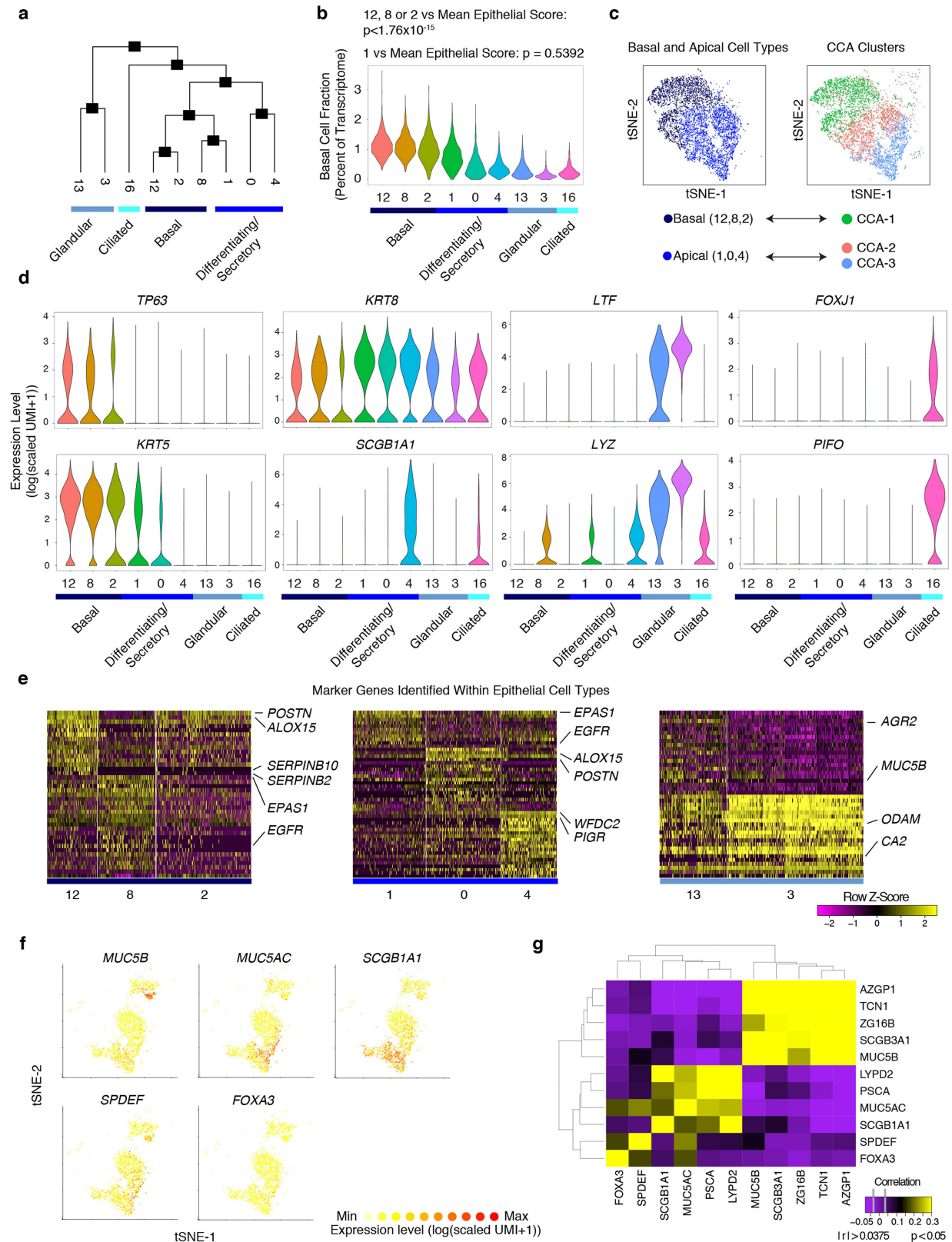
(see Supplementary Table 3 for full gene lists; genes identified by ROC test with AUC 0.691 for *CTGF*, 0.683 for *CXCL12*, 0.726 for *MYH11*); and a clustered correlation matrix of marker genes identified in single-cell data from fibroblasts. Note, clusters 4 and 5 are likely to represent doublets with epithelial cells and endothelial cells, respectively. Although we exclude these clusters from further formal analyses, we note that there may be interesting biology within pairs of cells that are found to interact more frequently than by chance. **c**, *t*-SNE plot of 1,143 endothelial cells ($n = 6$ non-polyp, $n = 6$ polyp samples), coloured by clusters identified through shared nearest neighbour (SNN) analysis (Supplementary Table 3; Methods), from CRS-EthSin; select marker gene overlays displaying count-based (UMI-collapsed) expression level ($\log(\text{scaled UMI} + 1)$) on a *t*-SNE plot (see Supplementary Table 3 for full gene lists; genes identified via ROC test with AUC 0.742 for *SELE*, 0.706 for *PODXL*, 0.822 for *PLAT*); and a clustered correlation matrix of marker genes identified in single-cell data from endothelial cells.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Mapping T2I mediators within EthSin non-polyp or polyp ecosystems and the identities of T cells. **a**, Dot plots of chemokines and lipid mediators with known roles in T2I mapped onto cell types divided by non-polyp or polyp disease state. Dot size represents fraction of cells within that type expressing the gene, and colour intensity represents binned ($\log(\text{scaled UMI} + 1)$) gene expression amongst expressing cells (related to Fig. 1d). **b**, Dot plot of inducers and effectors of T2I mapped onto cell types divided by non-polyp or polyp disease state. Dot size represents fraction of cells within that type expressing the gene, and colour intensity represents binned ($\log(\text{scaled UMI} + 1)$) gene expression amongst expressing cells (related to Fig. 1d). **c**, *t*-SNE plot of re-clustered T cells with select gene overlays displaying binned count-

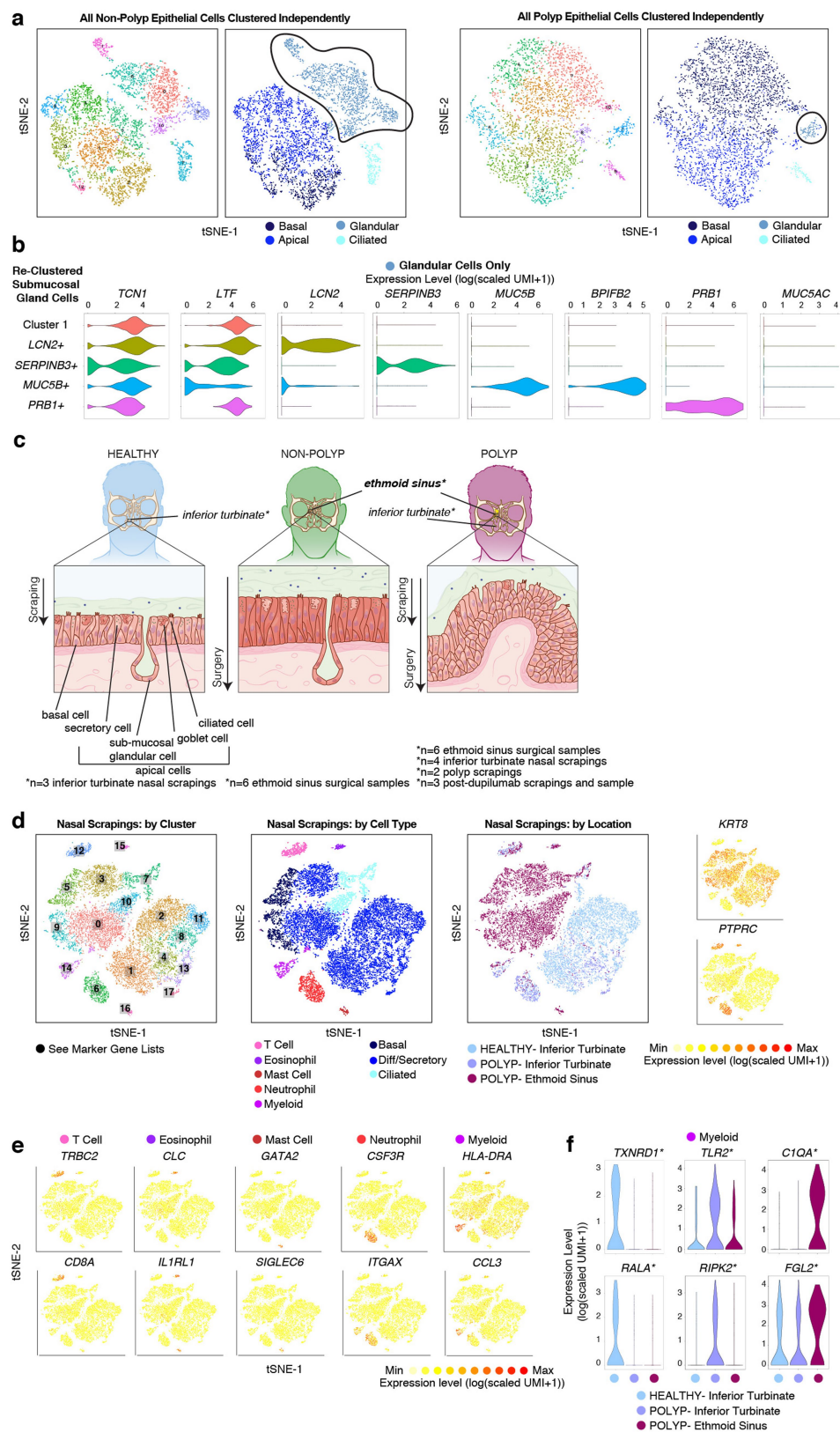
based expression level ($\log(\text{scaled UMI} + 1)$) for Th2A-specific genes (top row) and canonical T cell markers (bottom row); 835 T cells from $n = 6$ non-polyp and $n = 6$ polyp samples. **d**, Violin plot of five identified T cell clusters scored for expression of T cell receptor complex genes (for example, *TRAC* and *CD3E*, see Methods, Supplementary Table 4). Dots represent individual cells; 835 total T cells. **e**, Dot plot of inducers and effectors of Type 1 immunity across all cell types (note that *IL17F* was not detected). **f**, Dot plot of select GWAS risk alleles⁴¹ for allergic disease, mapped onto cell types. Dot size represents fraction of cells within that type expressing the gene, and colour intensity represents binned ($\log(\text{scaled UMI} + 1)$) gene expression amongst expressing cells (related to Fig. 1d).



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Relationship of EthSin epithelial cell clusters and secretory/glandular distinctions. **a**, A phylogenetic tree based on the average cell from each cluster of epithelial cell clusters in gene-space. **b**, Violin plot of expression contribution to a cell's transcriptome of basal cell genes (see Methods and Supplementary Table 4) across all epithelial cells. Cluster 12, 794 cells; cluster 8, 924 cells; cluster 2, 1,504 cells; cluster 1, 1,561 cells; cluster 0, 1,600 cells; cluster 4, 1,201 cells; cluster 13, 725 cells; cluster 3, 1,467 cells; cluster 16, 498 cells; Mann–Whitney *U*-test, with Bonferroni correction, $P < 1.76 \times 10^{-15}$, cluster 12, cluster 8 or cluster 2 versus the mean score of basal/apical epithelial cells; $P = 0.5392$, cluster 1 versus the mean score. **c**, Canonical correlation analysis (CCA) displaying our cell type annotations for basal and apical cells derived through clustering and biological curation alongside CCA clusters in *t*-

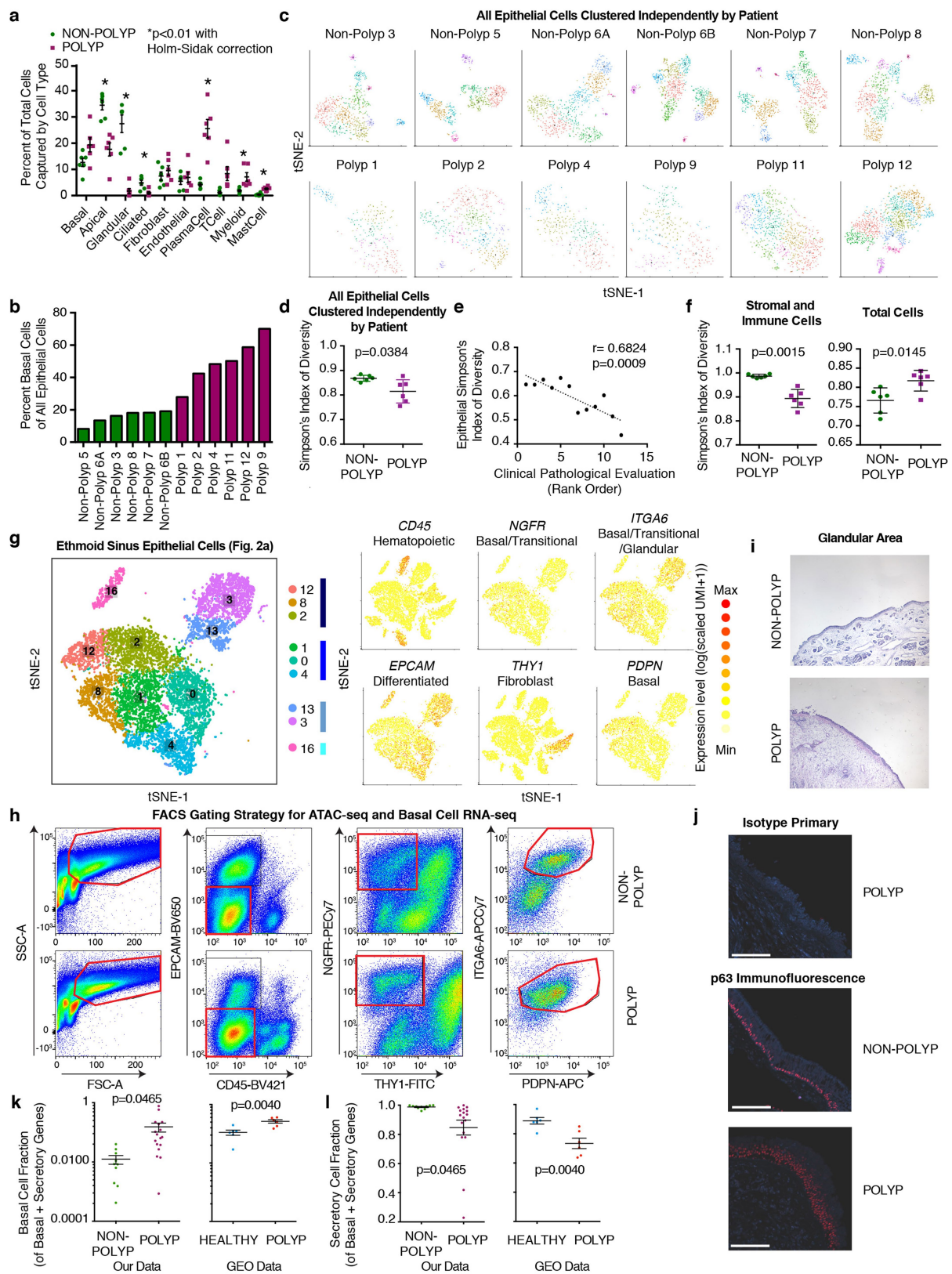
SNE space; 7,584 basal and apical cells. **d**, Violin plots for the count-based expression level ($\log(\text{scaled UMI} + 1)$) of selected marker genes for each identified epithelial cell subset; cell numbers as in **b**. **e**, Row-normalized heat map of the top marker genes identified by ROC test ($\text{AUC} > 0.6$) within each cell type for each cell cluster with genes displayed on *y* axis and cluster annotations on *x* axis (see Supplementary Table 3 for full gene lists). **f**, Select overlays on clusters 0 and 4 (differentiating/secretory) and 13 (glandular) displaying binned count-based expression level ($\log(\text{scaled UMI} + 1)$) in *t*-SNE space for canonical goblet (*MUC5B*, *MUC5AC*, *SPDEF*, *FOXA3*) and secretory (*SCGB1A1*) genes; 3,526 cells. **g**, A clustered correlation matrix of glandular, goblet, and secretory cell genes. Pearson's $\text{abs}(r) > 0.038$ is significant ($P < 0.05$) based on asymptotic *P* values.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Glandular cell subsets, their relationship to apical secretory cells, and immune cells recovered through nasal scrapings. **a**, *t*-SNE plots of 5,928 single epithelial cells ($n = 6$ non-polyp samples) and 4,346 single epithelial cells ($n = 6$ polyp samples) coloured by clusters identified through (left) shared nearest neighbour (SNN) analysis and (right) original biological curation of cell types (Supplementary Table 3; Methods) as illustrated in Fig. 2a. Note, cluster colours in left panels of each disease are not comparable but curated clusters in panels are, and glandular cells are highlighted for subsetting in next panel. **b**, Violin plots for the count-based expression level ($\log(\text{scaled UMI} + 1)$) of selected marker genes identified through marker discovery (ROC test) for each subset of glandular cells; 2,114 total cells (cluster 1, 791 cells; *LCN2* cluster, 709 cells; *SERPINB3* cluster, 283 cells; *MUC5B* cluster, 209 cells; *PRB1* cluster, 183 cells) with representation of every non-polyp patient in each cluster of cells (for example, no cluster is unique to one patient) and AUC metric 0.800 for *LCN2*, 0.736 for *SERPINB3*, 0.985 for *MUC5B*, 0.973 for *BPIFB2*, and 0.908 for *PRB1*. **c**, Samples were acquired through the two distinct methods of nasal scraping and ethmoid sinus surgical intervention. This allowed for sampling of healthy tissue from InfTurb (scraping, top left), CRS-EthSin-non-polyp tissue (surgery, top middle), CRS-EthSin-polyp tissue (surgery, top right), InfTurb of polyp-bearing individuals (scraping, top right) and CRS-EthSin-polyp tissue accessible for scraping (scraping, top right). Bottom panels, anatomy of the nasal turbinates (healthy and CRS polyp) and ethmoid sinus (CRS

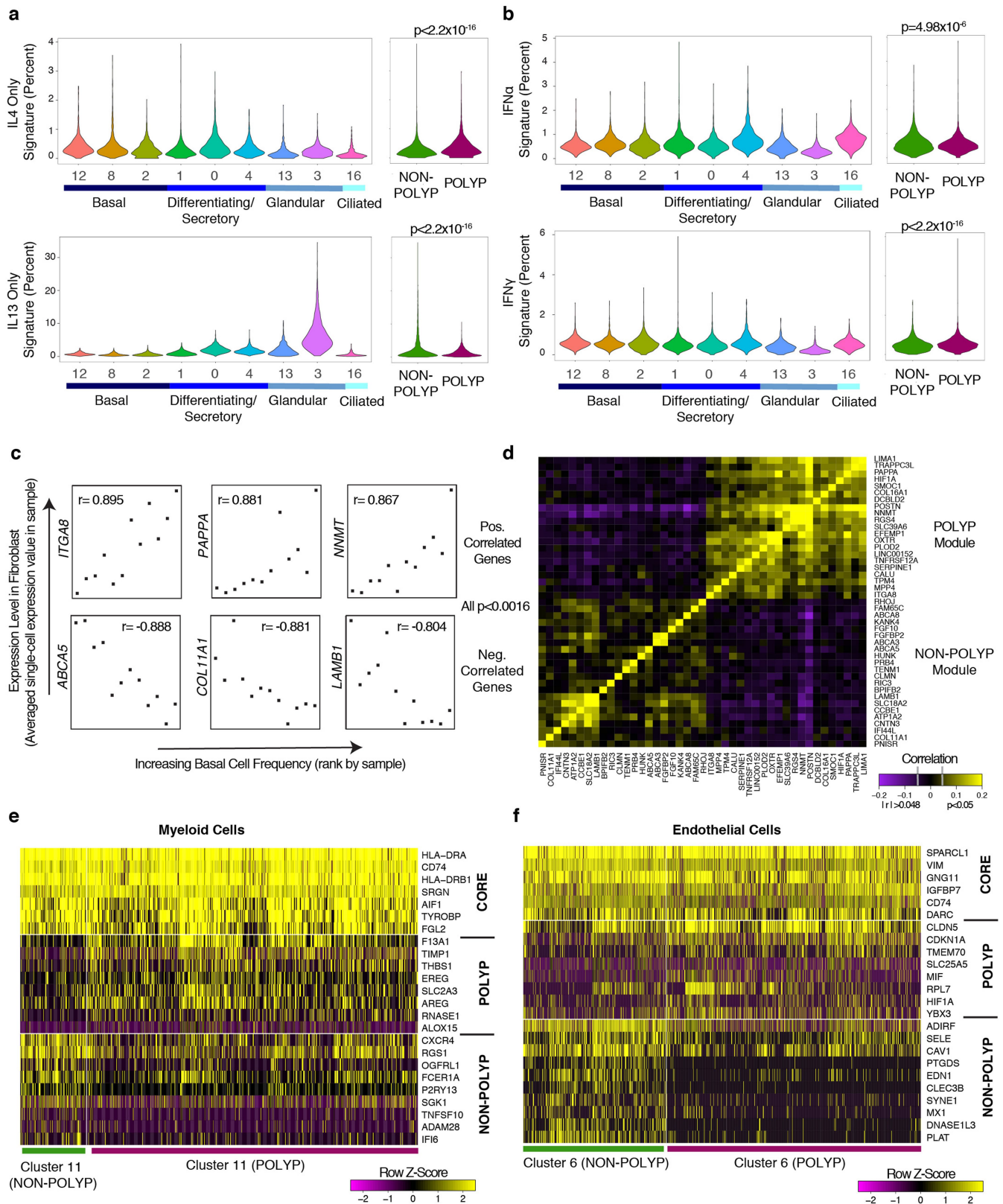
non-polyp and CRS polyp) where samples were acquired, highlighting the depth of cells recovered from each site related to Fig. 2. Healthy tissue is annotated with basal and apical cell types, including sub-mucosal glands. **d**, Left, *t*-SNE plot of 18,704 single cells from nasal scrapings ($n = 9$ samples) coloured by clusters identified through shared nearest neighbour (SNN) analysis (Supplementary Table 3; Methods). Middle, *t*-SNE plot coloured by cell types identified through marker discovery (ROC test) and biological curation of identified clusters (Supplementary Table 3; Methods). Right, *t*-SNE plot coloured by disease and tissue of origin from healthy InfTurb (7,603 cells; $n = 3$ samples), polyp-bearing patient InfTurb (2,298 cells; $n = 4$ samples) and polyp scraping directly from EthSin-polyp (8,803 cells; $n = 2$ samples), with adjacent select marker gene overlays displaying count-based UMI-collapsed expression level ($\log(\text{scaled UMI} + 1)$) for apical epithelial (*KRT8*) and haematopoietic (*PTPRC*) genes. **e**, Select marker gene overlays displaying count-based UMI-collapsed expression level ($\log(\text{scaled UMI} + 1)$) on a *t*-SNE plot from **a** for key cell types identified (see Supplementary Table 3 for full gene lists); area under the curve (AUC) 0.946 to 0.705 for all markers displayed. **f**, Violin plots for the count-based expression level ($\log(\text{scaled UMI} + 1)$) for key differentially expressed genes using ROC test within myeloid cells across disease states and tissues identified (Methods); 137 cells, $n = 3$ healthy inferior turbinate; 157 cells, $n = 4$ polyp inferior turbinate; 210 cells, $n = 2$ polyp ethmoid sinus samples; AUC 0.67 for *TXNRD1*, 0.615 for *RALA*, 0.647 for *TLR2*, 0.619 for *RIPK2*, 0.747 for *C1QA*, 0.674 for *FGL2*.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Changes in cellular composition between EthSin-non-polyp and EthSin-polyp tissue by scRNA-seq and flow cytometric gating and histological strategy for quantification and isolation of basal cells. **a**, The frequency of each cell type recovered amongst all cells within each patient sample ($n = 6$ non-polyp, $n = 6$ polyp) grouped by disease state. Two-sided t -test; apical, $P = 0.0003$; glandular, $P < 0.0001$; ciliated, $P = 0.0047$; plasma cell, $P = 0.00014$; myeloid, $P = 0.0098$; mast cell, $P = 0.00018$; all non-polyp versus polyp with Holm–Sidak correction for multiple comparisons. Data are mean \pm s.e.m. **b**, The frequency of basal cells amongst epithelial cells captured in scRNA-seq data displayed for each sample and coloured by non-polyp or polyp designation. **c**, t -SNE plots with each patient's cells clustered independently over a common list of most variable genes identified from all epithelial cells and with clustering parameters set constant to 12 principal components and resolution set to 1.4; minimum 789 cells in each plot; see Extended Data Fig. 1b and Supplementary Table 3 for specific cell numbers. **d**, Simpson's index of diversity, an indication of the total richness present within an ecosystem, over epithelial cell clusters identified in **c**, calculated for each patient; $n = 6$ non-polyp and $n = 6$ polyp samples. Two-tailed t -test, $P = 0.0384$. Data are mean \pm s.e.m. **e**, Correlation of Simpson's index of diversity calculated over epithelial cells against the ranked order of samples based on clinical pathological evaluation; $n = 6$ non-polyp and $n = 6$ polyp samples; $r = 0.6824$, $P = 0.009$. **f**, Simpson's index of diversity over stromal and immune cell types and total cells, calculated for each sample ($n = 6$ non-polyp and $n = 6$ polyp). Points represent individual

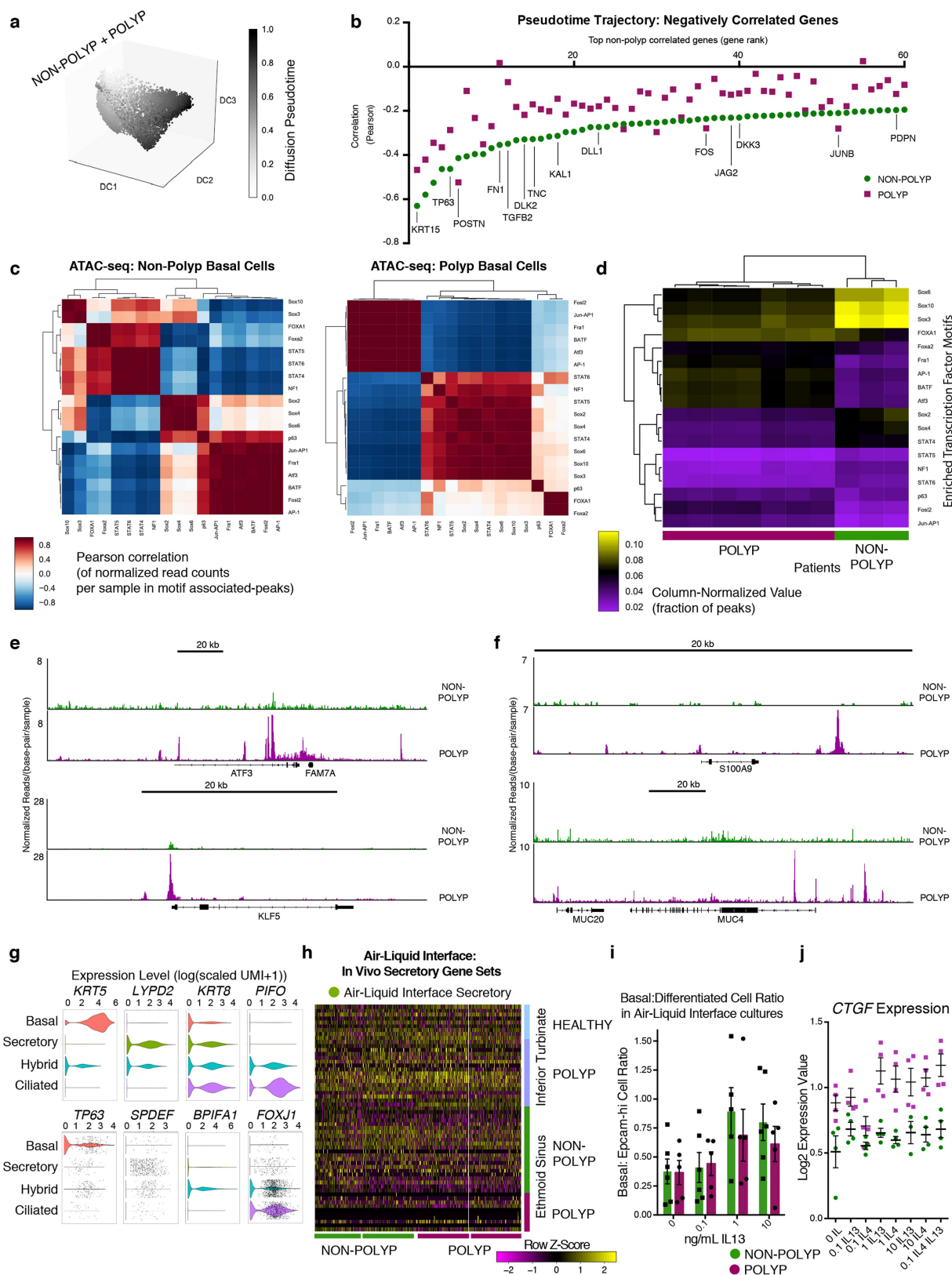
samples. Two-tailed t -test, $P = 0.0015$ (stromal and immune), $P = 0.0145$ (total cells), non-polyp versus polyp. Data are mean \pm s.e.m. **g**, Reproduced from Fig. 2a: t -SNE plot of 10,274 epithelial cells, coloured by clusters identified through SNN, with adjacent colour bars representing related cell clusters, and overlays displaying binned count-based expression level ($\log(\text{scaled UMI} + 1)$) of selected genes used to negatively (*CD45*, *EPCAM*, *THY1*) and positively (*NGFR*, *ITGA6*, *PDPN*) identify basal cells. **h**, Full flow cytometric gating strategy for quantification and isolation of basal cells from non-polyp and polyp tissue (related to Fig. 3c). **i**, Representative histology ($5\times$ magnification) of the glandular area detected in haematoxylin and eosin stained tissue sections from non-polyp or polyp patients; quantification in Fig. 3e. **j**, Representative immunofluorescence of p63⁺ cells (basal cell marker) relative to isotype control; quantification in Fig. 3d. Scale bar, 100 μm . **k**, Basal cell fraction of transcripts from bulk tissue RNA-seq data of our own dataset (related to Fig. 3g, h) and two GEO datasets containing healthy and healthy/polyp nasal mucosa biopsies. Our data: $n = 10$ non-polyp samples, $n = 17$ polyp samples. Reference data: $n = 6$ healthy, $n = 6$ polyp samples. Two-tailed t -test, $P = 0.0465$ (our data) and $P = 0.0040$ (GEO data). Data are mean \pm s.e.m. **l**, Secretory cell fraction of transcripts from bulk tissue RNA-seq data of our own dataset (related to Fig. 3g, h) and two GEO datasets containing healthy and healthy/polyp nasal mucosa biopsies. Our data, $n = 10$ non-polyp samples, $n = 17$ polyp samples; reference data, $n = 6$ healthy, $n = 6$ polyp samples. Two-tailed t -test, $P = 0.0465$ (our data) and $P = 0.0040$ (GEO data). Data are mean \pm s.e.m.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Epithelial cytokine signatures from CRS-EthSin tissue demonstrate T2I pattern, discovery of gene modules in the fibroblast niche which correlate with basal cell hyperplasia, and differential expression within myeloid and endothelial cells by polyp status. **a**, Violin plots of IL-4- or IL-13-uniquely induced gene signatures in respiratory epithelial cell clusters or grouped by disease state presented as expression contribution to a cell's transcriptome (see Methods, Fig. 4b for shared genes, and Supplementary Table 4). Cluster 12, 794 cells; cluster 8, 924 cells; cluster 2, 1,504 cells; cluster 1, 1,561 cells; cluster 0, 1,600 cells; cluster 4, 1,201 cells; cluster 13, 725 cells; cluster 3, 1,467 cells; cluster 16, 498 cells. Mann-Whitney U -test, $P < 2.2 \times 10^{-16}$, 0.305 IL-4 effect size (polyp versus non-polyp) and -0.448 IL-13 effect size (polyp versus non-polyp). **b**, Violin plots of IFN- α - or IFN- γ -induced gene signatures in respiratory epithelial cell clusters or grouped by disease state presented as expression contribution to a cell's transcriptome (see Methods, and Supplementary Table 4); cell numbers as in **a**. Mann-Whitney U -test, $P = 4.98 \times 10^{-6}$, -0.156 IFN- α effect size (polyp versus non-polyp). Mann-Whitney U -test, $P < 2.2 \times 10^{-16}$, 0.161 IFN- γ effect size (polyp versus non-polyp). **c**, Selected genes detected in fibroblasts from single-cell data which correlate with the samples ranked by basal cell frequency detected within each ecosystem. Non-polyp, $n = 6$; polyp, $n = 6$. All genes used: Spearman correlation, $\text{abs}(r) > 0.7651$, $P < 0.0037$. To determine

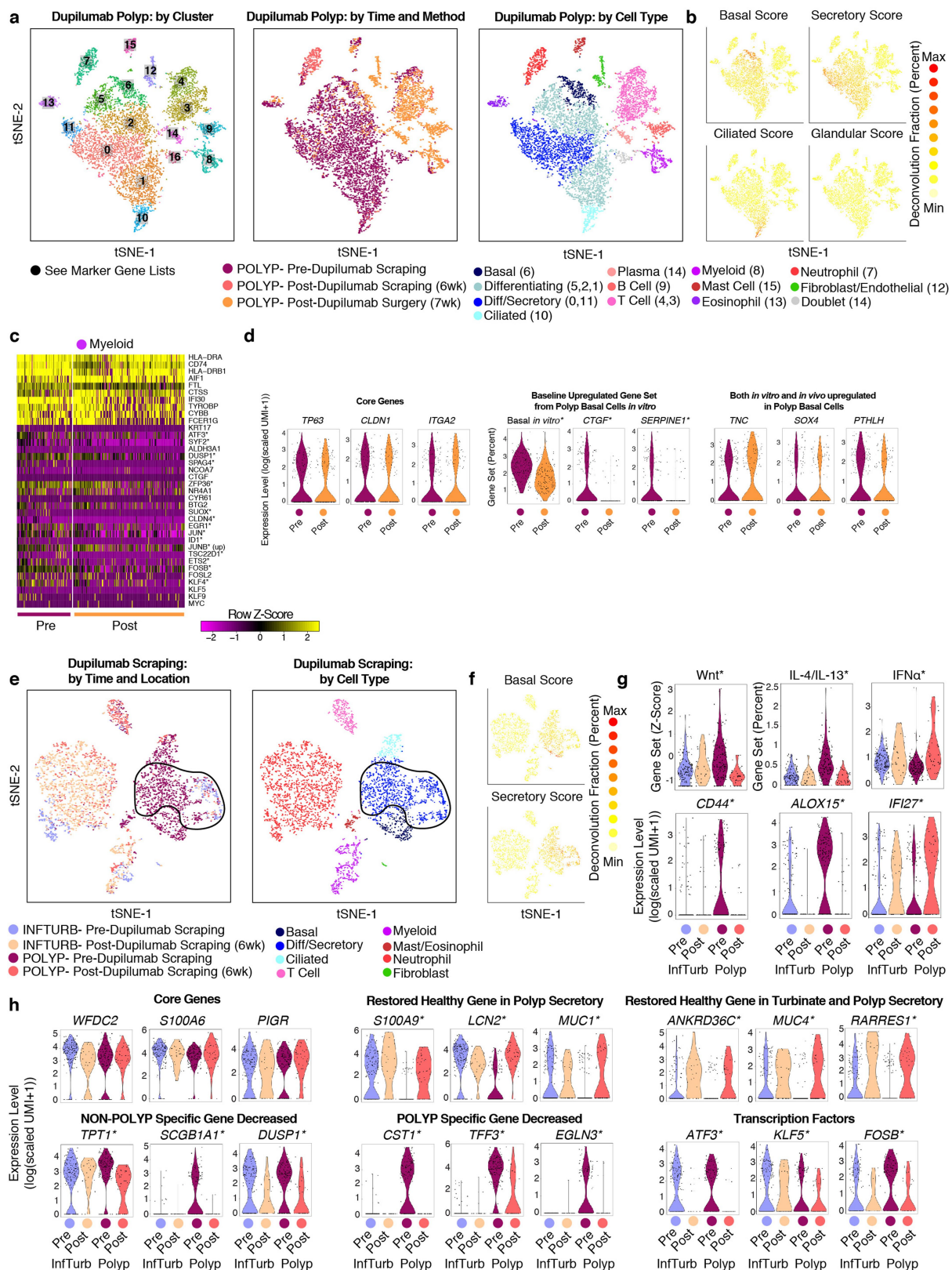
genes correlated in specific cell types (for example, fibroblasts) with the frequency of basal cells present in a cellular ecosystem, we correlated the average log-normalized single-cell count data for each gene to the rank of samples determined by increasing frequency of basal cells in each ecosystem (8.2% to 19.1% for non-polyp and 27.9% to 70.1% for polyp samples, Extended Data Fig. 7b). **d**, A clustered correlation matrix of genes identified as per **c** in single-cell data from fibroblasts; Pearson's $\text{abs}(r) > 0.048$ is significant ($P < 0.05$) based on asymptotic P values. **e**, Row-normalized heat map for myeloid cells from ethmoid sinus with select genes displayed on the y axis, including a core myeloid signature (ROC test myeloid cells versus rest of cells, AUC > 0.8), and genes differentially expressed (bimodal test) by disease state, with disease-state annotations on x axis. Bimodal test, all non-core genes $P < 0.0002$ or less with Bonferroni correction for multiple hypothesis testing based on number of genes tested. **f**, Row-normalized heat map for endothelial cells from ethmoid sinus with select genes displayed on y axis including a core basal signature (ROC test endothelial cells versus rest of cells, AUC > 0.75), and genes differentially expressed (bimodal test) by disease state, with disease-state annotations on x axis. Bimodal test, all non-core genes $P < 2.43 \times 10^{-6}$ or less with Bonferroni correction for multiple hypothesis testing based on number of genes tested.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Pseudotime analysis on basal and differentiating/secretory cell clusters from EthSin, transcriptional motif enrichments in non-polyp and polyp basal cells, and the identity of cell types in air-liquid interface cultures. **a**, Pseudotime analysis using diffusion mapping (see Methods) of selected clusters of epithelial cells, here displaying diffusion pseudotime (related to Fig. 4d). Clusters 8/1/4, 3,516 cells; clusters 12/2/0, 4,064 cells. $n = 6$ non-polyp, $n = 6$ polyp samples. Diffusion map and diffusion coefficients (DC) are calculated over the set of basal and apical marker genes identified in Fig. 1a (see Supplementary Table 3). **b**, The top 60 negatively correlated genes expressed in non-polyp cells with pseudotime trajectory and Pearson correlation values for genes in polyp cells also displayed; differential correlation coefficient analysis using Fisher's z -statistic, accounting for number of cells in each group (specific genes highlighted, all $>2z$; full results including Bonferroni corrected P values in Supplementary Table 3). **c**, Correlation matrices (row and column clustered) of the normalized read counts per sample in motif-associated peaks for non-polyp or polyp samples. Pearson correlation, $n = 3$ non-polyp, $n = 7$ polyp. **d**, A column-normalized heat map (row and column clustered) for the fraction of peaks with a motif corresponding to accessibility of the respective transcription factor displayed by patient. $n = 3$ non-polyp, $n = 7$ polyp. **e**, IGV tracks for *ATF3* and *KLF5* based on peaks detected and averaged by non-polyp and polyp samples from ATAC-seq profiling. **f**, IGV tracks for *S100A9* and *MUC4* based on peaks detected and averaged by non-polyp and polyp samples from ATAC-seq profiling.

g, Violin plots for the count-based expression level ($\log(\text{scaled UMI} + 1)$) for key marker genes using ROC test across cell types identified in Fig. 5a, Supplementary Table 3. 1,345 basal; 6,420 secretory; 6,381 hybrid; and 2,027 ciliated cells from $n = 2$ non-polyp and $n = 2$ polyp patients. AUC = 0.943 (*KRT5*), 0.667 (*TP63*), 0.644 (*LYPD2*), <0.55 (*SPDEF*), <0.55 (*KRT8*), 0.602 (*BPIFA1*), 0.813 (*PIFO*), 0.73 (*FOXJ1*). **h**, Row-normalized heat map for ALI secretory cells (subsampling to 300 cells per donor) as in Fig. 2f of the top in vivo secretory marker genes identified by ROC test (AUC > 0.662) with select genes displayed on y axis including a core secretory signature (ROC test, secretory cells versus rest of cells), and then within secretory cells, ROC test was used to identify marker genes within each disease/location category; and basal-cell derived annotations on x axis (see Supplementary Table 3 for full gene lists, all AUC > 0.65 for markers displayed in Fig. 2f). **i**, Quantification of flow cytometry for the ratio of basal to Epcam^{hi} cells (gating as in Extended Data Fig. 7h) from ALI cultures at 21 days, stimulated with the indicated doses of IL-13. Points represent individual biological replicates; $n = 6$ non-polyp, $n = 5$ polyp samples for each dose. Two-way ANOVA; not significant between disease groups at any dose tested; Two-way ANOVA, $P = 0.0224$ for IL-13 dose. Data are mean \pm s.e.m. **j**, Expression levels for *CTGF* (\log_2 expression value of log-normalized count data) in basal cells from non-polyp or polyp individuals across doses of cytokines displayed. $n = 4$ samples each dose. Two-way ANOVA $P < 0.0260$ for *CTGF*; all conditions non-polyp versus polyp except 0.1 ng ml^{-1} IL-4 dose for *CTGF*.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | In vivo blockade with an anti-IL-4R α monoclonal antibody shifts secretory cell state towards healthy-associated genes. **a**, Left, *t*-SNE plot of 8,764 single cells (related to Fig. 5e) from the nasal polyps of an individual treated with dupilumab (IL-4R α monoclonal antibody) (1 patient, sampled at $n = 3$ time points), coloured by clusters identified through SNN analysis (Supplementary Table 3; Methods); middle, *t*-SNE plot coloured by time point and tissue of origin from polyp pre-dupilumab scraping (5,731 cells), from polyp post-dupilumab scraping (647 cells), and polyp post-dupilumab surgical sample (2,386 cells); right, *t*-SNE plot coloured by cell types identified through marker discovery (ROC test) and biological curation of identified clusters (Supplementary Table 3; Methods). **b**, Select cell-type specific score overlays for cell types indicated in original core dataset (see Supplementary Table 3 for full gene list). **c**, Row-normalized heat map for myeloid cells of the top marker genes identified by ROC test (AUC > 0.8) with select genes displayed on *y* axis including a core myeloid signature (ROC test myeloid cells versus rest of cells), and then genes found to be differentially expressed from Fig. 5f in basal cells, and treatment annotations on *x* axis. Bimodal test. *, differential genes in both basal cells and myeloid cells pre- versus post-treatment ($P < 0.003$ or less with Bonferroni correction for multiple hypothesis testing based on number of genes tested). **d**, Violin plots for basal cells (200 cells pre-dupilumab and 151 cells post-dupilumab, noted in **a**) for the count-based expression level ($\log(\text{scaled UMI} + 1)$), except where indicated for gene scores, fraction of transcriptome and *z*-score (see Methods, Supplementary Table 4 for gene set used) for key basal cell genes for selected biological processes, or from the baseline upregulated gene set from polyp basal cells in vitro (Fig. 5c). Differential expression testing for decreased expression post-treatment using bimodal test not significant except where indicated (* $P < 0.00087$ or less with Bonferroni correction for multiple hypothesis testing based on number of genes tested; see Supplementary Table 3 for full list). Basal in vitro score pre versus post: two-tailed *t*-test, $P < 3.897 \times 10^{-15}$, effect size 0.822. **e**, *t*-SNE plot of 4,486 single cells (related to Figs. 2e, 5e) from the inferior turbinate or nasal polyps of an anti-IL-4R α (dupilumab) treated individual ($n = 4$ samples) coloured by time point and tissue of origin (left) from inferior turbinate pre-dupilumab scraping (643

cells), from inferior turbinate post-dupilumab scraping (1,596 cells), polyp pre-dupilumab scraping (1,600 cells), and polyp post-dupilumab scraping (647 cells). *t*-SNE plot coloured by cell types (right) identified through marker discovery (ROC test) and biological curation of identified clusters (Supplementary Table 3; Methods); black outline indicates cells considered in **g**. **f**, Select deconvolution score overlays for cell types indicated in original core dataset (see Supplementary Table 3 for full gene list). **g**, Violin plot for the gene set score over Wnt pathway (*z*-score) and expression contribution to a cell's transcriptome over IFN- α - and IL-4/IL-13-commonly induced gene signature in secretory cells grouped as in **e** and sub-sampled to a maximum of 150 cells from each disease or location category from inferior turbinate pre-dupilumab scraping (150 cells), from inferior turbinate post-dupilumab scraping (23 cells), polyp pre-dupilumab scraping (150 cells) and polyp post-dupilumab scraping (38 cells) (see Methods, Supplementary Table 3, Supplementary Table 4 for gene lists used). Two-tailed *t*-test; Wnt score pre versus post polyp tissue, effect size 1.02, $P = 1.091 \times 10^{-14}$; Wnt score pre versus post inferior turbinate tissue, effect size -0.17 , $P = 0.3706$; IL-4/IL-13 score pre versus post polyp tissue, effect size 1.17, $P < 2.2 \times 10^{-16}$; IL-4/IL-13 score pre versus post inferior turbinate tissue, effect size -0.17 , $P = 0.163$; IFN- α score pre versus post polyp tissue, effect size -1.25 , $P = 4.254 \times 10^{-5}$; IFN- α score pre versus post inferior turbinate tissue, effect size -0.304 , $P = 0.2766$. * $P < 7.81 \times 10^{-6}$ or less between pre- and post-treated polyp, differential expression testing for decreased expression post-treatment using bimodal test. **h**, Violin plots of secretory cells grouped as in **e** and sub-sampled to a maximum of 150 cells from each disease or location category from inferior turbinate pre-dupilumab scraping (150 cells), inferior turbinate post-dupilumab scraping (23 cells), polyp pre-dupilumab scraping (150 cells) and polyp post-dupilumab scraping (38 cells) for the count-based expression level ($\log(\text{scaled UMI} + 1)$) and for secretory cell genes from the gene set used in Fig. 2f affected by treatment within anatomical regions indicated by heading. * $P < 6.36 \times 10^{-5}$ or less except *KLF5* ($P = 0.0033$) and *FOSB* ($P = 0.0053$), differential expression testing for decreased expression post-treatment using bimodal test with Bonferroni correction for multiple hypothesis testing based on number of genes tested, see Supplementary Table 3 for all genes tested.

RAP2 mediates mechanoresponses of the Hippo pathway

Zhipeng Meng¹, Yunjiang Qiu^{2,3}, Kimberly C. Lin¹, Aditya Kumar⁴, Jesse K. Placone⁴, Cao Fang¹, Kuei-Chun Wang^{4,5}, Shicong Lu¹, Margaret Pan¹, Audrey W. Hong¹, Toshiro Moroishi^{1,6,7}, Min Luo^{1,8}, Steven W. Plouffe¹, Yarui Diao², Zhen Ye², Hyun Woo Park^{1,9}, Xiaoqiong Wang¹⁰, Fa-Xing Yu¹¹, Shu Chien^{4,5}, Cun-Yu Wang¹², Bing Ren^{2,13}, Adam J. Engler⁴ & Kun-Liang Guan^{1*}

Mammalian cells are surrounded by neighbouring cells and extracellular matrix (ECM), which provide cells with structural support and mechanical cues that influence diverse biological processes¹. The Hippo pathway effectors YAP (also known as YAP1) and TAZ (also known as WWTR1) are regulated by mechanical cues and mediate cellular responses to ECM stiffness^{2,3}. Here we identified the Ras-related GTPase RAP2 as a key intracellular signal transducer that relays ECM rigidity signals to control mechanosensitive cellular activities through YAP and TAZ. RAP2 is activated by low ECM stiffness, and deletion of RAP2 blocks the regulation of YAP and TAZ by stiffness signals and promotes aberrant cell growth. Mechanistically, matrix stiffness acts through phospholipase C γ 1 (PLC γ 1) to influence levels of phosphatidylinositol 4,5-bisphosphate and phosphatidic acid, which activates RAP2 through PDZGEF1 and PDZGEF2 (also known as RAPGEF2 and RAPGEF6). At low stiffness, active RAP2 binds to and stimulates MAP4K4, MAP4K6, MAP4K7 and ARHGAP29, resulting in activation of LATS1 and LATS2 and inhibition of YAP and TAZ. RAP2, YAP and TAZ have pivotal roles in mechanoregulated transcription, as deletion of YAP and TAZ abolishes the ECM stiffness-responsive transcriptome. Our findings show that RAP2 is a molecular switch in mechanotransduction, thereby defining a mechanosignalling pathway from ECM stiffness to the nucleus.

YAP and TAZ function as essential effectors of mechanotransduction to regulate cell proliferation and differentiation^{3–7}. When cells are shifted from stiff to soft matrices, YAP and TAZ translocate from the nucleus to the cytoplasm, and are thereby inactivated. However, it is unclear how ECM stiffness is signalled to the Hippo pathway. Because small GTPases function as molecular switches in many biological processes⁸, we searched for small GTPases that affect localization of YAP and TAZ in cells seeded on soft (1 kPa) or stiff (40 kPa) matrices (Supplementary Information). RAP2A was identified because its overexpression induced cytoplasmic translocation of YAP and TAZ even on a stiff matrix (Fig. 1a). No other GTPases, including the closely related RAP1B, H-RAS, K-RAS, and N-RAS, showed similar activity (Extended Data Fig. 1a).

At high stiffness, both wild-type MCF10A cells and those in which RAP2A, RAP2B and RAP2C were deleted (RAP2-KO MCF10A cells) showed nuclear localization of YAP and TAZ (Fig. 1b, c). At low stiffness, wild-type cells exhibited mainly cytoplasmic YAP and TAZ, whereas RAP2-KO MCF10A cells retained YAP and TAZ in the nucleus (Fig. 1c). Deletion of RAP2 in HEK293A cells also suppressed cytoplasmic translocation of YAP and TAZ induced by low stiffness (Fig. 1d, e,

Extended Data Fig. 1b). Expression of the YAP and TAZ target genes *CTGF*, *CYR61*, and *ANKRD1* was repressed by low stiffness in wild-type cells, but not in RAP2-KO cells (Fig. 1f). Similar results were observed in human mesenchymal stem cells (Extended Data Fig. 1c–e), in which RAP2 deletion suppressed differentiation into adipocytes (Extended Data Fig. 1f, g). In luminal breast cancer MCF7 cells, ECM stiffness modulated localization of YAP and TAZ in a RAP2-dependent manner, whereas the basal type MDA-MB-468 cells showed constitutively cytoplasmic localization of YAP and TAZ regardless of stiffness (Extended Data Fig. 1h–l). TWIST1 and β -catenin have been reported to show nuclear–cytoplasmic shuttling in response to physical cues^{9,10}. TWIST1, but not β -catenin, displayed nuclear–cytoplasmic translocation in response to ECM stiffness (Extended Data Fig. 2a). However, RAP2 deletion had no obvious effect on TWIST1 localization.

The activity of small GTPases is switched on and off by binding of GTP and GDP, respectively. A RalGDS-RBD pull-down assay showed that low stiffness promotes binding of GTP to RAP2 (Fig. 2a, Extended Data Fig. 2b). Unlike wild-type RAP2A, the GTP-binding-deficient mutant RAP2A(S17N) did not induce cytoplasmic translocation of YAP and TAZ (Extended Data Fig. 2b, c). PDZGEF1 and PDZGEF2 are RAP2 activators^{11–13}, and the interaction of RAP2 with PDZGEF1 was enhanced by low stiffness (Extended Data Fig. 2d). We generated cells lacking both activators (PDZGEF1/2-dKO cells; Extended Data Fig. 2e, f) and discovered that they were defective in cytoplasmic translocation of YAP and TAZ (Fig. 2b, c) and target gene repression (Extended Data Fig. 2g) in response to low stiffness. Deletion of PDZGEF1 and PDZGEF2 blunted activation of RAP2 by low stiffness (Fig. 2d), and PDZGEF1 overexpression induced cytoplasmic translocation of YAP and TAZ in wild-type but not RAP2-KO cells (Extended Data Fig. 2h, i).

Phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P₂) activates RAP2 through PDZGEF1 and PDZGEF2 at the plasma membrane after PtdIns(4,5)P₂ is converted into phosphatidic acid (PA) by phospholipase D1 (PLD1) and PLD2¹³. Using a GFP-tagged PtdIns(4,5)P₂ reporter, we observed that PtdIns(4,5)P₂ was enriched at the plasma membrane at low stiffness (Fig. 2e, Extended Data Fig. 2j). Focal adhesions decrease PtdIns(4,5)P₂ by activating PLC γ 1^{14,15}. Inhibition of PLC γ 1 by U73122 induced PtdIns(4,5)P₂ accumulation (Extended Data Fig. 2k). We hypothesized that ECM stiffness regulates RAP2, YAP and TAZ by modulating focal adhesion and local PtdIns(4,5)P₂ abundance at the plasma membrane. The focal adhesion kinase (FAK) inhibitor PF573228 or the PLC γ 1 inhibitor U73122 increased binding of GTP by RAP2 and cytoplasmic translocation of YAP and TAZ (Extended Data

¹Department of Pharmacology and Moores Cancer Center, University of California San Diego, La Jolla, CA, USA. ²Ludwig Institute for Cancer Research, La Jolla, CA, USA. ³Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA. ⁴Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. ⁵Institute of Engineering in Medicine, University of California San Diego, La Jolla, CA, USA. ⁶Department of Molecular Enzymology, Faculty of Life Sciences, Kumamoto University, Kumamoto, Japan. ⁷Center for Metabolic Regulation of Healthy Aging, Faculty of Life Sciences, Kumamoto University, Kumamoto, Japan. ⁸State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, Sichuan, China. ⁹Department of Biochemistry, College of Life Science & Biotechnology, Yonsei University, Seoul, South Korea. ¹⁰Robert J. Tomisch Pathology & Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA. ¹¹Children's Hospital and Institutes of Biomedical Sciences, Fudan University, Shanghai, China. ¹²Division of Oral Biology and Medicine, School of Dentistry, University of California Los Angeles, Los Angeles, CA, USA. ¹³Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA. *e-mail: kuguan@ucsd.edu

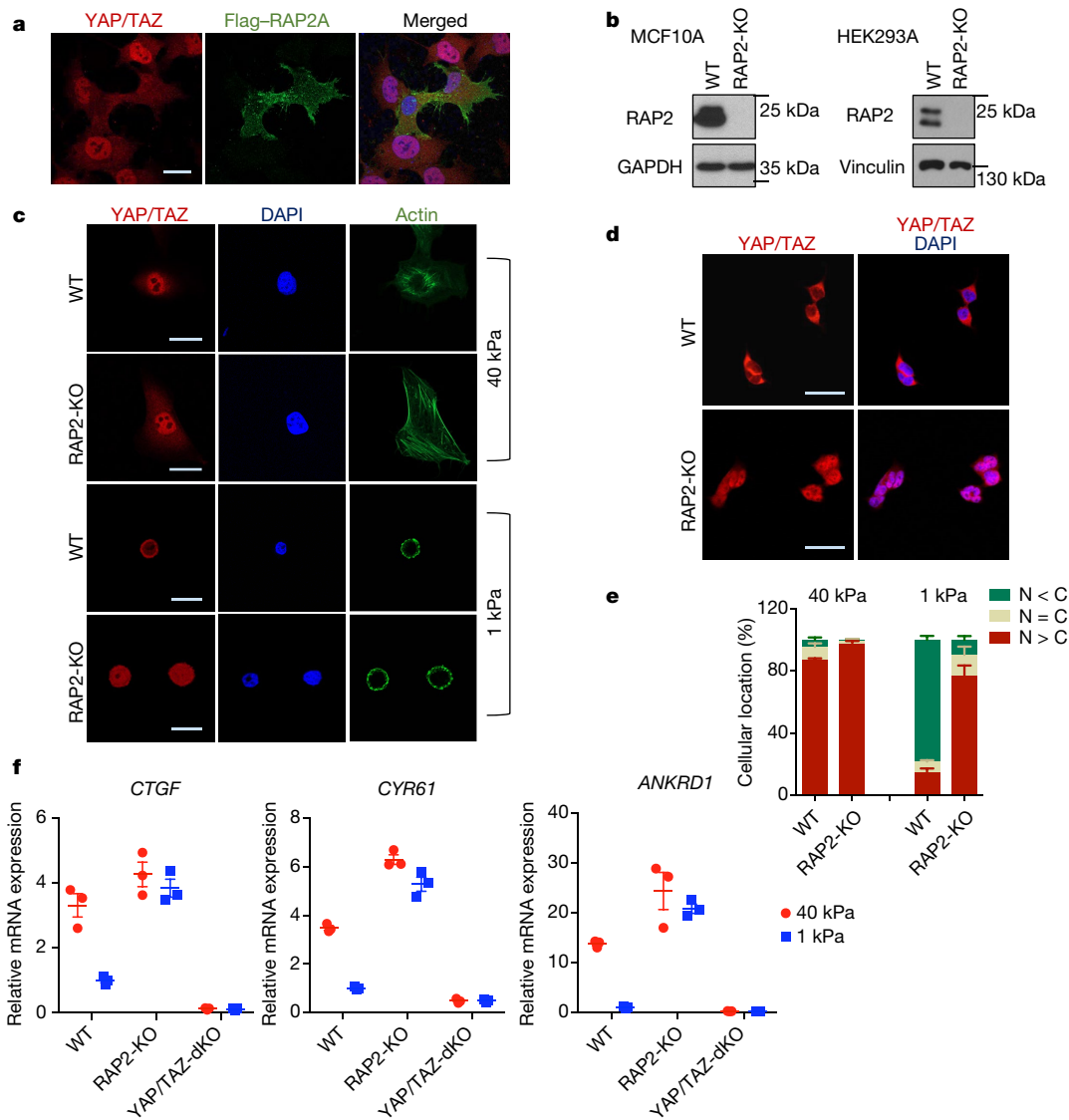


Fig. 1 | RAP2 mediates regulation of YAP and TAZ by ECM stiffness.

a, Overexpression of Flag–RAP2A induces cytoplasmic translocation of YAP and TAZ in HEK293A cells on a stiff (40 kPa) matrix. Merged, combined signals from YAP and TAZ (red), Flag (green), and DAPI (blue). **b**, Immunoblots showing deletion of RAP2A, RAP2B and RAP2C (RAP2-KO) in MCF10A and HEK293A cells. WT, wild-type. **c**, Immunofluorescence showing that RAP2-KO MCF10A cells, unlike wild-type cells, retain nuclear YAP and TAZ at low stiffness (1 kPa). The experiments in **b** and **c** were repeated independently twice with similar results. **d**, Deletion of RAP2A, RAP2B and RAP2C in HEK293A cells

blocks cytoplasmic localization of YAP and TAZ by low stiffness.

e, Quantification of YAP and TAZ localization, presented as mean \pm s.e.m., in HEK293A cells. $N < C$, less YAP and TAZ in nucleus than in cytoplasm; $N = C$, similar levels of YAP and TAZ in cytoplasm and nucleus; $N > C$, more YAP and TAZ in nucleus than in cytoplasm. **f**, RAP2 is required for regulation of YAP and TAZ target genes *CTGF*, *CYR61*, and *ANKRD1* by stiffness in HEK293A cells. Data are presented as mean \pm s.e.m. For **e** and **f**, $n = 3$ biologically independent samples. Scale bars, 25 μ m.

Fig. 3a–c). By contrast, the PLD1 and PLD2 inhibitor BML279 reduced GTP binding by RAP2 and induced the accumulation of YAP and TAZ in the nucleus (Extended Data Fig. 3d–f). The effects of PtdIns(4,5) P_2 on localization of YAP and TAZ were confirmed by experiments involving knockdown of PLC γ 1 and combined knockdown of PLD1 and PLD2 (Extended Data Fig. 3g–i).

The nuclear–cytoplasmic shuttling of YAP and TAZ is generally controlled by LATS1- and LATS2-dependent phosphorylation². However, the role of the Hippo kinase cascade in the regulation of YAP and TAZ by mechanotransduction is not clear^{4–7}. We found that low ECM stiffness induced phosphorylation of LATS1 and LATS2, and YAP and TAZ, in wild-type cells, and that this phosphorylation was substantially blunted in RAP2-KO cells (Fig. 3a). Furthermore, RAP2 induced YAP phosphorylation in a GTP-binding-dependent manner (Extended Data Fig. 3m). We proposed that RAP2 controls localization of YAP and TAZ via the Hippo pathway. Consistent with this notion, deletion of

LATS1 and LATS2 or combined deletion of MST1, MST2 and MAP4Ks abolished regulation of YAP and TAZ by ECM rigidity (Extended Data Fig. 3n, o). Hippo pathway core components were similarly required for RAP2 to induce phosphorylation and cytoplasmic translocation of YAP and TAZ (Extended Data Fig. 4a, b). The role of LATS1 and LATS2 in this regulation was confirmed in LATS1/2-dKO mouse embryonic fibroblasts, as well as in NF2-KO or MOB1A/MOB1B-dKO HEK293A cells^{16,17} (Extended Data Fig. 4c, d).

MAP4K4, TNIK (MAP4K7), and ARHGAP29 are RAP2 effectors^{18–20}. Notably, ARHGAP29 is one of the RhoGAPs that are transcriptionally activated by YAP^{21,22}. MAP4K4 kinase activity was stimulated by low stiffness in wild-type but not RAP2-KO cells (Extended Data Fig. 4e, f). Moreover, low stiffness induced MAP4K4 phosphorylation, as indicated by reduced mobility, in a RAP2-dependent manner (Extended Data Fig. 4g, h). Deletion of the RAP2-interacting citron domain¹⁹ in MAP4K4 abolished its regulation by RAP2 (Extended Data Fig. 4i, j), and the

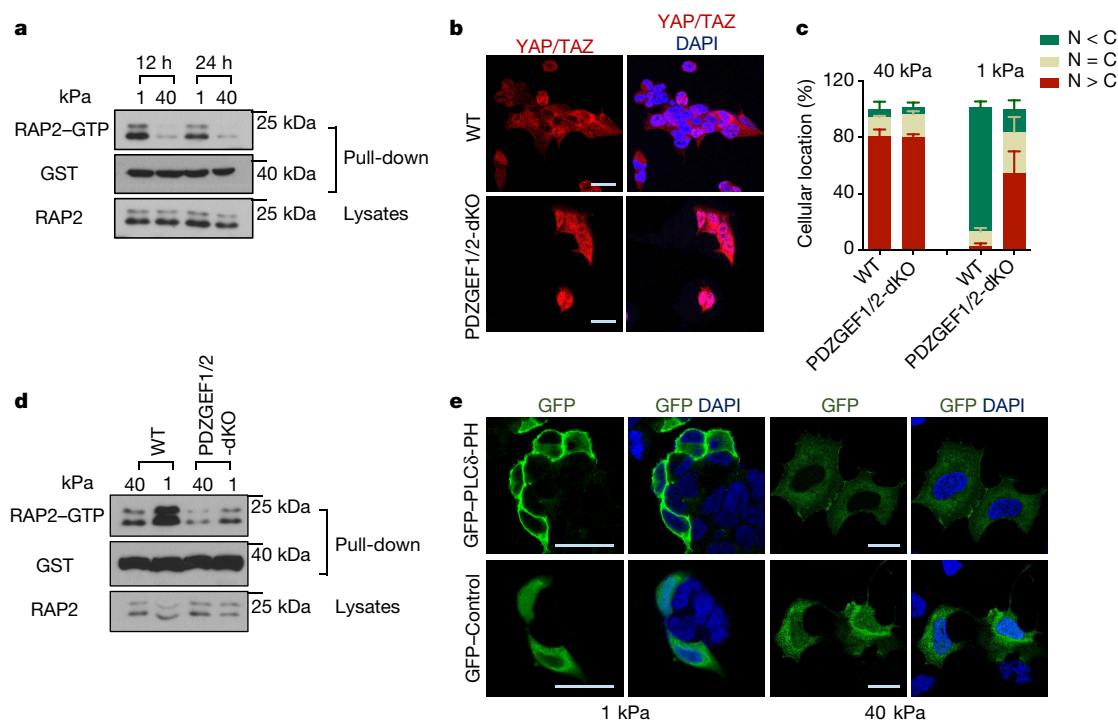


Fig. 2 | ECM stiffness acts via PDZGEF1 and PDZGEF2 to regulate RAP2. **a**, RAP2 is activated by low stiffness. Pull-down of GTP-bound RAP2 from cells at 1 kPa and 40 kPa using GST–RalGDS–RBD. **b**, Deletion of PDZGEF1 and PDZGEF2 compromises translocation of YAP and TAZ at 1 kPa. **c**, Quantification of YAP and TAZ localization, presented as mean \pm s.e.m., in **b**. $n = 3$ (40 kPa) or 4 (1 kPa) biologically

independent samples. **d**, PDZGEF1 and PDZGEF2 mediate regulation of RAP2 by stiffness. Experiments were similar to **a**. **e**, High stiffness reduces enrichment of PtdIns(4,5) P_2 at the plasma membrane. A GFP-tagged PtdIns(4,5) P_2 reporter PLC δ -PH domain was transfected into cells and detected with anti-GFP antibodies. The experiments in **a**, **d**, **e** were repeated independently twice with similar results. Scale bars, 25 μ m.

mutant also failed to rescue the YAP and TAZ translocation defect in cells lacking MST1, MST2 and the MAP4Ks (MST1/2–MAP4K1/2/3/4/6/7–8KO cells) (Fig. 3b). Notably, a recent study showed that the *Drosophila* MAP4K4/6/7 homologue Msn regulates Yki in response to tension²³. In addition, overexpression of RAP2A led to inactivation of RhoA (Extended Data Fig. 5a), a potent activator for YAP and TAZ^{7,16} (Extended Data Fig. 5b–d). This action of RAP2A on RhoA was mediated by ARHGAP29, because YAP phosphorylation induced by ARHGAP29 required its Rho-GAP domain and the Hippo kinase cascade (Fig. 3c, Extended Data Fig. 5e). Deletion of ARHGAP29 compromised inactivation of RhoA by low stiffness (Extended Data Fig. 5f, g). Therefore, RAP2 acts through MAP4K4, MAP4K6, MAP4K7 and ARHGAP29 to inhibit YAP and TAZ. Consistent with this finding, MAP4K4/6/7–ARHGAP29–4KO cells were resistant to RAP2-induced cytoplasmic translocation and phosphorylation of YAP and TAZ (Fig. 3d, Extended Data Fig. 5h), and displayed impaired responses to ECM stiffness (Fig. 3e, Extended Data Fig. 5i). Collectively, our data reveal a signalling axis that links matrix stiffness to regulation of YAP and TAZ as follows: focal adhesion \rightarrow PLC γ 1 \rightarrow PtdIns(4,5) P_2 \rightarrow PA \rightarrow PDZGEF \rightarrow RAP2 \rightarrow ARHGAP29 and MAP4K \rightarrow LATs (Fig. 3f), which works in parallel to the cell spreading–RhoA–cytoskeleton tension-mediated YAP and TAZ translocation mechanism proposed previously^{4,7}.

RAP2 is activated by cell–cell contact¹², which also presents a mechanical cue to cells and inhibits YAP and TAZ^{2,17}. Deletion of RAP2 moderately increased nuclear YAP and TAZ at high confluence (Extended Data Fig. 6a, b). Combined deletion of RAP2 with MST1 and MST2 resulted in stronger nuclear accumulation and gene transactivation of YAP and TAZ (Extended Data Fig. 6a–e). Deletion of RAP2, MST1 and MST2 is required to blunt phosphorylation of LATs and inactivation of YAP and TAZ, suggesting that confluency signalling is complex and additional routes, such as cellular junctions, contribute to activation of LATs1 and LATs2.

Deletion of RAP2 selectively enhanced cell growth only at low stiffness (Extended Data Fig. 7a). To assess the role of RAP2 in tumorigenesis, we performed three assays: acinus formation, anchorage-independent

growth, and xenotransplantation. First, we used a 3D-culture system with low stiffness hydrogels to assay the formation of MCF10A acini⁹ (Extended Data Fig. 7b); aberrant acinus formation represents irregular cell growth and malignant transformation^{9,24}. Wild-type cells formed normal acini whereas RAP2-KO cells generated multi-acinar structures (Fig. 4a, b, Extended Data Fig. 7c), and knockdown of YAP and TAZ significantly reduced the development of aberrant acini. Because MST1 and MST2 mediate some physical signals independent of RAP2 (Extended Data Fig. 6), we generated MCF10A cells lacking RAP2A, RAP2B, RAP2C, MST1 and MST2 (RAP2–MST1/2–KO; Extended Data Fig. 7d). Whereas MST1/2–dKO cells formed relatively normal acini, the RAP2–MST1/2–KO cells formed large acini with invasive behaviours (Extended Data Fig. 7e, f) even at 150 Pa, similar to the stiffness of normal breast tissue (Extended Data Fig. 7g, h). Second, a colony-formation assay in soft agar showed that RAP2–MST1/2–KO cells displayed anchorage-independent growth (Extended Data Fig. 8a). Third, RAP2–MST1/2–KO cells showed substantial xenograft growth in immune-deficient mice, whereas MST1/2–dKO cells did not (Fig. 4c, Extended Data Fig. 8b, c). RAP2–MST1/2–KO xenografts contained abundant MCF10A cells recapitulating the acinus and duct formation of breast tissue, whereas MST1/2–dKO xenografts consisted of mainly host cells with a small number of MCF10A cells (Extended Data Fig. 8d, e). Moreover, the RAP2–MST1/2–KO cells displayed architectural and cytological atypia with signs of malignancy (Extended Data Fig. 8e). Consistently, knockdown of YAP and TAZ suppressed xenograft growth (Fig. 4d, Extended Data Fig. 8f–h). The function of RAP2 in stiffness-regulated growth was confirmed in a xenograft model using H-RAS–V12-expressing MCF10A cells^{25,26} (Fig. 4e, f, Extended Data Fig. 8i, j). We used LOX-overexpressing fibroblasts or semisynthetic hyaluronan-derived hydrogels (soft: 0.40 ± 0.03 kPa; stiff: 8.98 ± 0.33 kPa)^{27,28} to assess the effect of stiffness on cell growth in xenograft models (Extended Data Fig. 9). Under low stiffness or with control fibroblasts, RAP2-KO cells grew significantly larger xenografts than wild-type cells, whereas the growth advantage of RAP2-KO cells was decreased at high stiffness or in the presence of LOX-overexpressing fibroblasts (Fig. 4g, h, Extended Data Fig. 9).

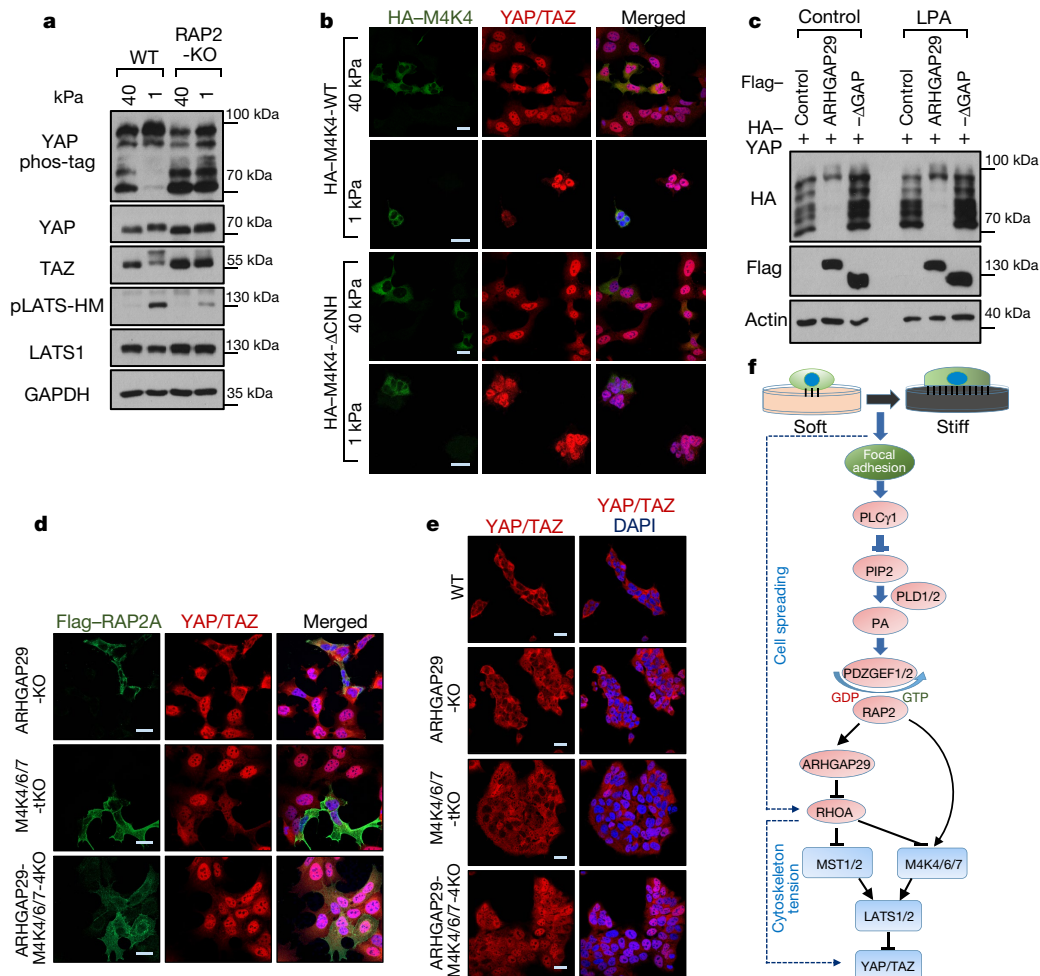


Fig. 3 | RAP2 inhibits YAP and TAZ through MAP4K4, MAP4K6, MAP4K7 and ARHGAP29. **a**, RAP2 is important for stiffness-regulated phosphorylation of LATS1, LATS2, YAP and TAZ. Phos-tag gel detects YAP phosphorylation by mobility shift. pLATS-HM detects phosphorylation of LATS1 and LATS2 in the hydrophobic motif. **b**, Expression of wild-type MAP4K4 (M4K4-WT), but not the citron domain-deleted mutant (M4K4-ΔCNH), rescued localization of YAP and TAZ in MM-8KO (MST1/2-MAP4K1/2/3/4/6/7-8KO) cells at low stiffness. HA, haemagglutinin. Merged, combined signals from HA-MAP4K4 (green), YAP and TAZ (red), and DAPI (blue). **c**, Overexpression of wild-type ARHGAP29, but not the GAP domain-deleted mutant (ARHGAP29-ΔGAP), induced phosphorylation of YAP.

YAP and TAZ are transcriptional co-activators that generate functional output through gene transcription². We performed RNA sequencing (RNA-seq) with RAP2-KO, LATS1/2-dKO, and YAP/TAZ-dKO HEK293A cells to investigate their transcriptional responses to ECM stiffness (Fig. 4i, Extended Data Fig. 10a). In wild-type cells, low stiffness led to downregulation of 814 genes and upregulation of 513 genes. These genes are involved in metabolic processes, such as RNA and macromolecule biosynthesis, and morphogenesis (Supplementary Information). YAP and TAZ are ‘gate-keepers’ that are responsible for almost all the stiffness-responsive genes, as deletion of YAP and TAZ or LATS1 and LATS2 abolished most of the changes in expression of these genes (Fig. 4i), including *AMOTL2* and *LGR5* (Extended Data Fig. 10b). Consistent with its role in mechanosignalling, RAP2 deletion completely abolished the changes in expression of 40–50% (not including those partially blunted) of stiffness-responsive genes (Fig. 4i).

To assess the interplay of RAP2 and Hippo pathway components in stiffness-dependent gene regulation, we enriched YAP- and TAZ-activating genes that were downregulated by YAP and TAZ knockout at high stiffness and upregulated by LATS1 and LATS2 knockout at low

stiffness, and YAP- and TAZ-repressing genes that were conversely regulated (Extended Data Fig. 10c, d). These genes define an ECM-Hippo transcriptome that comprises nearly a third of the genes affected by ECM stiffness (Extended Data Fig. 10e, f). RAP2 deletion completely abolished the responses to ECM in about 50% of ECM-Hippo transcriptome genes (Extended Data Fig. 10g, h), and also partially compromised many genes. The RAP2-regulated ECM-Hippo transcriptome revealed that RAP2 controls genes involved in cell growth and adhesion (for example, *CTGF*, *CYR61*), as well as morphogenesis and development (for example, *KRT8*, *KRT18*, *GDF6*), through the Hippo pathway to respond to ECM stiffness (Fig. 4j, Supplementary Information).

This study shows that RAP2 is an intracellular mechanotransducer that relays extracellular mechanical signals to transcriptional regulation through the Hippo pathway. ECM stiffness acts through RAP2 and its downstream Hippo kinase cascade to modulate a YAP- and TAZ-mediated mechanoresponsive transcriptome. The identification of this signalling axis provides mechanistic insights into how cellular machinery is driven by mechanical stimuli.

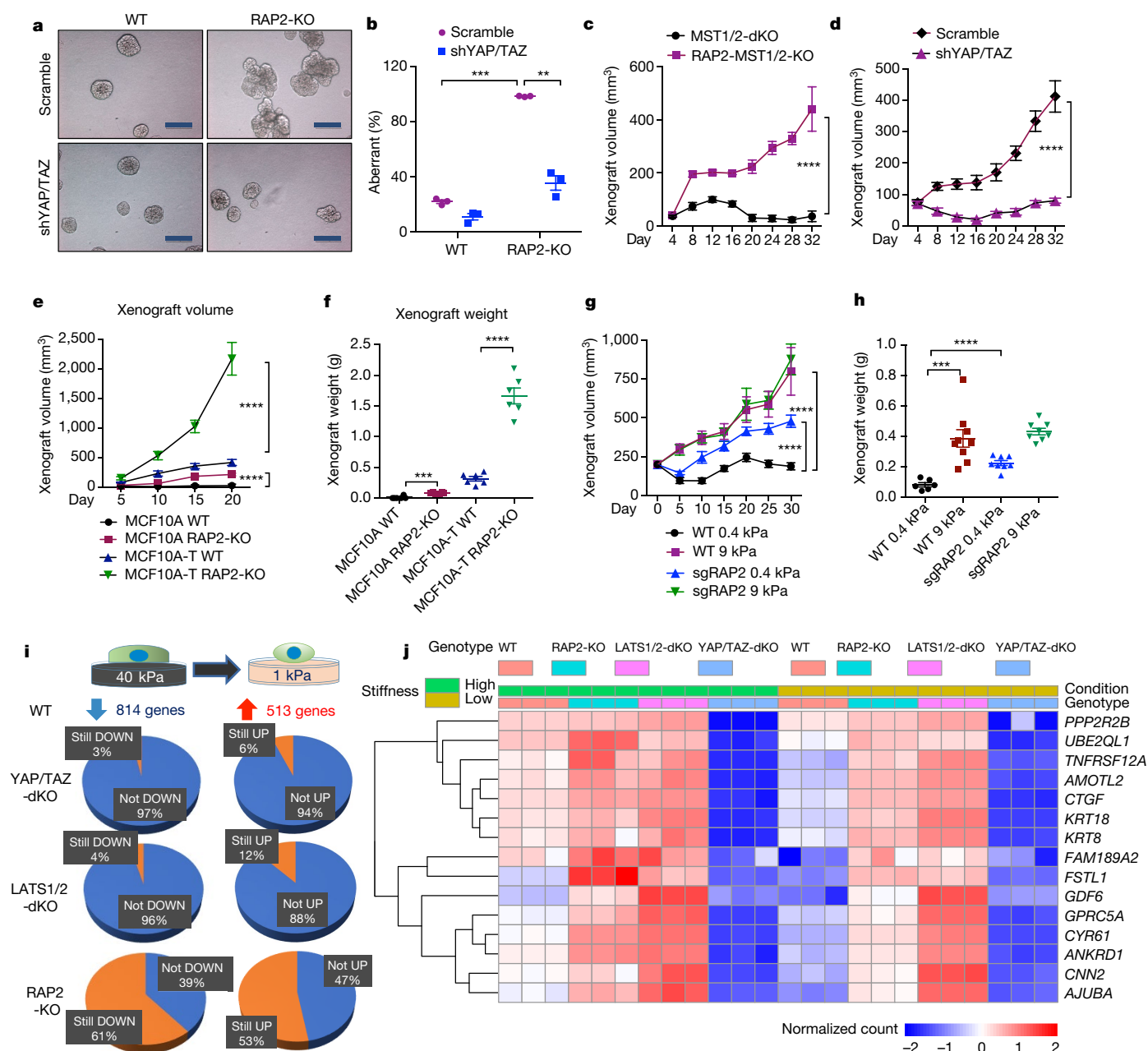


Fig. 4 | RAP2 suppresses cell transformation and regulates the ECM stiffness transcriptome through Hippo and YAP. **a**, Knockdown of YAP and TAZ (shYAP/TAZ) suppresses the aberrant acinus growth caused by deletion of RAP2A, RAP2B and RAP2C in MCF10A cells. Scale bars, 100 μ m. Scramble, control shRNA. **b**, Quantification of aberrant acini from three hydrogels (mean \pm s.e.m.). Two-tailed *t*-test: ***RAP2-KO versus wild-type, $P = 0.00022$; **shYAP/TAZ versus scramble, $P = 0.0067$. **c**, MST1/2-dKO and RAP2-MST1/2-KO MCF10A cells were injected subcutaneously into NOD/SCID mice. The tumour volume is shown as mean \pm s.e.m. Two-way ANOVA, $n = 6$ biologically independent xenografts, **** $P < 0.0001$. **d**, RAP2-MST1/2-KO MCF10A cells with scramble or shYAP/TAZ were injected into NOD/SCID mice. Tumour volume presented as mean \pm s.e.m. Two-way ANOVA, $n = 6$ biologically independent xenografts, **** $P < 0.0001$. **e**, MCF10A and MCF10A-T cells with RAP2 deleted were injected into nude mice.

The tumour volume is presented as mean \pm s.e.m. Two-way ANOVA, $n = 6$ biologically independent xenografts, **** $P < 0.0001$. **f**, Tumour weights on day 20 from **e** (mean \pm s.e.m.). Two-tailed *t*-test: *** $n = 6$, $P = 0.0001$; **** $n = 6$, $P = 0.00007$. **g**, MCF7 cells with lentivirus-mediated CRISPR deletion of RAP2 (sgRAP2) embedded in 200 μ l of 0.4 or 9.0 kPa hyaluronan-based gel were injected into nude mice. The xenograft volume is shown as mean \pm s.e.m. Two-way ANOVA test, $n = 6$ (wild-type, 0.4 kPa), 8 (sgRAP2, 0.4 or 9.0 kPa), or 9 (wild-type, 9.0 kPa) biologically independent xenografts, **** $P < 0.0001$. **h**, Tumour weights on day 31 from **g** (mean \pm s.e.m.). Two-tailed *t*-test: *** $n = 6$ or 9, $P = 0.0006$; **** $n = 6$ or 8, $P = 0.00002$. **i**, Expression of genes downregulated or upregulated by low stiffness in wild-type cells assessed by RNA-seq and compared with YAP/TAZ-dKO, LATS1/2-dKO, and RAP2-KO cells. **j**, Heat map of representative ECM-responsive genes regulated by RAP2 and the Hippo pathway.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0444-0>.

Received: 24 April 2017; Accepted: 12 July 2018;

Published online 22 August 2018.

- Humphrey, J. D., Dufresne, E. R. & Schwartz, M. A. Mechanotransduction and extracellular matrix homeostasis. *Nat. Rev. Mol. Cell Biol.* **15**, 802–812 (2014).
- Meng, Z., Moroishi, T. & Guan, K. L. Mechanisms of Hippo pathway regulation. *Genes Dev.* **30**, 1–17 (2016).
- Halder, G., Dupont, S. & Piccolo, S. Transduction of mechanical and cytoskeletal cues by YAP and TAZ. *Nat. Rev. Mol. Cell Biol.* **13**, 591–600 (2012).
- Aragona, M. et al. A mechanical checkpoint controls multicellular growth through YAP/TAZ regulation by actin-processing factors. *Cell* **154**, 1047–1059 (2013).

5. Codelia, V. A., Sun, G. & Irvine, K. D. Regulation of YAP by mechanical strain through Jnk and Hippo signaling. *Curr. Biol.* **24**, 2012–2017 (2014).
6. Wada, K., Itoga, K., Okano, T., Yonemura, S. & Sasaki, H. Hippo pathway regulation by cell morphology and stress fibers. *Development* **138**, 3907–3914 (2011).
7. Dupont, S. et al. Role of YAP/TAZ in mechanotransduction. *Nature* **474**, 179–183 (2011).
8. Cherfils, J. & Zeghouf, M. Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiol. Rev.* **93**, 269–309 (2013).
9. Wei, S. C. et al. Matrix stiffness drives epithelial–mesenchymal transition and tumour metastasis through a TWIST1–G3BP2 mechanotransduction pathway. *Nat. Cell Biol.* **17**, 678–688 (2015).
10. Benham-Pyle, B. W., Pruitt, B. L. & Nelson, W. J. Mechanical strain induces E-cadherin-dependent Yap1 and β -catenin activation to drive cell cycle entry. *Science* **348**, 1024–1027 (2015).
11. de Rooij, J. et al. PDZ-GEF1, a guanine nucleotide exchange factor specific for Rap1 and Rap2. *J. Biol. Chem.* **274**, 38125–38130 (1999).
12. Monteiro, A. C. et al. Trans-dimerization of JAM-A regulates Rap2 and is mediated by a domain that is distinct from the *cis*-dimerization interface. *Mol. Biol. Cell* **25**, 1574–1585 (2014).
13. Gloerich, M. et al. Rap2A links intestinal cell polarity to brush border formation. *Nat. Cell Biol.* **14**, 793–801 (2012).
14. Carloni, V., Romanelli, R. G., Pinzani, M., Laffi, G. & Gentilini, P. Focal adhesion kinase and phospholipase C gamma involvement in adhesion and migration of human hepatic stellate cells. *Gastroenterology* **112**, 522–531 (1997).
15. Zhang, X. et al. Focal adhesion kinase promotes phospholipase C- γ 1 activity. *Proc. Natl Acad. Sci. USA* **96**, 9021–9026 (1999).
16. Plouffe, S. W. et al. Characterization of Hippo pathway components by gene inactivation. *Mol. Cell* **64**, 993–1008 (2016).
17. Meng, Z. et al. MAP4K family kinases act in parallel to MST1/2 to activate LATS1/2 in the Hippo pathway. *Nat. Commun.* **6**, 8357 (2015).
18. Myagmar, B. E. et al. PARG1, a protein-tyrosine phosphatase-associated RhoGAP, as a putative Rap2 effector. *Biochem. Biophys. Res. Commun.* **329**, 1046–1052 (2005).
19. Machida, N. et al. Mitogen-activated protein kinase kinase kinase 4 as a putative effector of Rap2 to activate the c-Jun N-terminal kinase. *J. Biol. Chem.* **279**, 15711–15714 (2004).
20. Taira, K. et al. The Traf2- and Nck-interacting kinase as a putative effector of Rap2 to regulate actin cytoskeleton. *J. Biol. Chem.* **279**, 49488–49496 (2004).
21. Qiao, Y. et al. YAP regulates actin dynamics through ARHGAP29 and promotes metastasis. *Cell Reports* **19**, 1495–1502 (2017).
22. Porazinski, S. et al. YAP is essential for tissue tension to ensure vertebrate 3D body shape. *Nature* **521**, 217–221 (2015).
23. Li, Q. et al. Ingestion of food particles regulates the mechanosensing misshapen–yorkie pathway in *Drosophila* intestinal growth. *Dev. Cell* **45**, 433–449 (2018).
24. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**, 256–268 (2003).
25. Yoh, K. E. et al. Repression of p63 and induction of EMT by mutant Ras in mammary epithelial cells. *Proc. Natl Acad. Sci. USA* **113**, E6107–E6116 (2016).
26. Dawson, P. J., Wolman, S. R., Tait, L., Heppner, G. H. & Miller, F. R. MCF10AT: a model for the evolution of cancer from proliferative breast disease. *Am. J. Pathol.* **148**, 313–319 (1996).
27. Serban, M. A., Scott, A. & Prestwich, G. D. Use of hyaluronan-derived hydrogels for three-dimensional cell culture and tumor xenografts. *Curr. Protoc. Cell Biol.* Ch. 10, Unit 10 14 (2008).
28. Levental, K. R. et al. Matrix crosslinking forces tumor progression by enhancing integrin signaling. *Cell* **139**, 891–906 (2009).

Acknowledgements K.C.L., A.W.H., and S.W.P. are supported by the T32 GM007752 training grant, A.K. by T32AR060712, and J.K.P. by F32HL126406. K.-L.G. is supported by grants from the NIH (CA196878, CA217642, GM51586, DE015964) as is A.J.E. (R21CA217735, R01CA206880). A.J.E. is also supported by NSF grant 1463689, and A.K. is supported by the NSF graduate research fellowship program and an ARCS/Roche Foundation Scholar Award in Life Science. H.W.P. is supported by KHIDI grant H17C1560.

Reviewer information *Nature* thanks M. Sudol, V. Weaver and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.M. and K.-L.G. conceived the project and wrote the manuscript. Z.M., K.C.L., C.F., S.L., M.P., T.M. and M.L. performed in vitro cell assays, CRISPR knockout, quantitative real-time PCR, immunofluorescence, and xenograft studies. A.K., J.K.P., K.-C.W., A.W.H., S.C. and A.J.E. assisted in manufacturing hydrogels and with immunofluorescence experiments. S.W.P. and H.W.P. provided knockout cell lines. Y.Q., Y.D., Z.Y. and B.R. performed the next-generation sequencing and bioinformatics analyses. X.W. performed the pathological analyses. F.-X.Y., C.-Y.W., B.R. and A.J.E. provided technical support.

Competing interests K.-L.G. is a co-founder of Vivace Therapeutics. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0444-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0444-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to K.-L.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cell culture. HEK293A cells were maintained in DMEM containing 10% fetal bovine serum. MCF10A cells were from ATCC and maintained as previously described²⁴. The pre-malignant derivative of MCF10A cells, MCF10A-T, were generated by infecting MCF10A cells with a lentiviral vector expressing a constitutively active H-Ras mutant (G12V) as previously described^{25,26}. Adipocyte-derived mesenchymal stem cells (MSCs) were cultured in HMSC growth medium from Cell Applications Inc., and were differentiated into adipocytes and stained according to a modified protocol²⁹. MCF7 and MDA-MB-468 cells were from ATCC and maintained in DMEM/F12 with 10% FBS. Insulin (0.01 mg/ml) was used for maintaining MCF7 cells. The cell lines were tested to be free of mycoplasma contamination and not experimentally authenticated.

Plasmids. Flag-pLJM1-RAP2A and pRK5-HA-GST-RAP2A plasmids were provided by D. M. Sabatini (Addgene 19311 and 14952). PDZGEF1 and ARHGAP29 coding sequences were subcloned from cDNA clones BC117321 and BC093741 (Transomics Technology), respectively. GFP-C1-PLCdelta-PH was a gift from T. Meyer (Addgene 21179).

The CRISPR-Cas9 system was used to delete genes in HEK293A, MCF10A, MSC, MCF7 and MDA-MB-468 cells. The plasmids px459 v2 and lentiCRISPR v2 were provided by F. Zhang (Addgene 62988 and 52961).

The single-guide RNA (sgRNA) sequences targeting individual genes were as following: *RAP2A* #1: GATGCGCGAGTACAAAGTGG; *RAP2A* #2: GTATTTC TCG ATG AAGGTGC; *RAP2B* #1: CATGAGAGAGTACAAAGTGG; *RAP2B* #2: GGAGCCCCGTCACGAAGTGC; *RAP2C* #1: GGTGAAGGTGAGACTCATGA; *RAP2C* #2: AGTGACAAAGTGCACAGTAA; *PDZGEF1* (also known as *RAPGEF2*) #1: CCCATAAGCTGAGTGTAG; *PDZGEF1* (also known as *RAPGEF2*) #2: CC AGCTAACCATGGAGTAT; *PDZGEF2* (also known as *RAPGEF6*) #1: TCAA CGCCTGCTAGCGCCA; *PDZGEF2* (also known as *RAPGEF6*) #2: GCCACCCG AGCGGACTCCCG; *ARHGAP29* #1: CTCTACTTACATATT TCAA; *ARHGAP29* #2: AGTTATTATATACGTCTAG.

RAP2-GTP and RhoA-GTP binding assay. The GST-RalGDS-RBD-expressing BL21 strain was provided by R. Firtel, and the GST-Rhotekin-RBD-expressing BL21 strain was a gift from J. Heller Brown at the University of California, San Diego. The recombinant proteins were purified and stored bound to glutathione-agarose beads. The binding of RAP2-GTP and RhoA-GTP from cell lysates to RalGDS-RBD-agarose and Rhotekin-RBD-agarose beads, respectively, was performed in a buffer containing 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 25 mM MgCl₂, 10% glycerol, 0.5% NP-40 substitute and 1 mM DTT.

Cell culture with 2D or 3D polyacrylamide-based hydrogels. 2D culture on hydrogels of high (40.40 ± 2.39 kPa) or low (1.00 ± 0.31 kPa) stiffness was as described elsewhere³⁰. In brief, 10 µg/ml human placenta fibronectin or 25–50 µg/ml rat tail collagen I were used to coat the sulfo-sanpah-activated hydrogels according to the preferences of the cell lines. 3D culture followed a protocol with the modification that the Matrigel base was replaced with fibronectin-coated hydrogels^{9,24}.

Staining and microscopy. For immunofluorescence, cells were fixed in 4% formaldehyde/PBS for 10 min and then were treated with 0.1% Triton X-100 or saponin (only for staining PIP2 GFP reporter) for 15 min. After blocking, the cells were stained with corresponding antibodies. Most images were captured with a Nikon Eclipse Ti confocal microscope and then were exported from NIS elements imaging software. Images in Fig. 2d and Extended Data Fig. 4c, d were taken with an Olympus FV1000 confocal microscope, and Image J was used to merge the signals from channels. For immunohistochemistry, xenografts were subjected to heat-induced antigen retrieval using 10 mM sodium citrate buffer followed by 3% H₂O₂ for 30 min to quench endogenous peroxidase activity. Sections were incubated overnight at 4°C with YAP/TAZ antibody and detected using Vectastain elite ABC kit and DAB Peroxidase Substrate kit (Vector Laboratories) according to the manufacturer's protocol.

Preparation of semi-synthetic hyaluronan-derived hydrogels. Under aseptic conditions, Glycosil (ESI Bio, GS222), Gelin-S, (ESI Bio, GS231), and Extralink (ESI Bio, GS3006) were dissolved in degassed water (ESI Bio, GS240) according to the manufacturer's directions. To make soft hydrogels, stock concentrations of 10 mg/ml Glycosil, 10 mg/ml Gelin-S, and 5 mg/ml Extralink were made per manufacturer's directions. To make stiff hydrogels, concentrated stocks of Glycosil and Extralink were prepared by solubilization in reduced volumes to make 2× Glycosil and 5× Extralink. Prior to use, aliquots were taken from each vial to make solutions at 1:5 ratios of Extralink: (Glycosil + Gelin-S) and 5× Extralink: (2× Glycosil + Gelin-S) for soft and stiff hydrogels, respectively. For all conditions, the amount of Gelin-S was kept constant to ensure the same number of gelatin-based cell binding sites.

In order to perform atomic force microscopy, the Glycosil and Gelin-S components were mixed thoroughly and then the Extralink was added to initiate gelation. Subsequently, 50 µl of the mixture was added drop-wise to DCMS-treated glass slides and a methacrylated coverslip was placed on top. Each sample was prepared

in triplicate. The slides were then incubated at 37°C until complete gelation. For the samples tested, this occurred within 30–40 min of incubation at 37°C. Upon gelation, hydrogel stiffness was measured by AFM. Hydrogels were then placed in PBS containing 1% antibiotic/antimycotic at 37°C and the stiffness was measured 1, 24, and 48 h after mixing.

AFM measurement procedure. AFM was performed to measure hydrogel stiffness as previously described³¹. In brief, indentations were made using a pyrex-nitride probe with a pyramid tip (spring constant ~0.04 N/m, 35° half-angle opening, NanoAndMore USA Corporation, PNP-TR) connected to a MFP-3D Bio Atomic Force Microscope (Oxford Instruments) mounted on a Ti-U fluorescence inverted microscope (Nikon Instruments). Probes were calibrated using the Igor 6.34A software (WaveMetrics). Samples were then loaded on the AFM, submersed in PBS, and indented at a velocity of 2 µm/s with a trigger force of 2 nN. About 20 force measurements were performed over a 90 µm × 90 µm region per gel. Measurements were made each day for three separate gels per condition. Elastic modulus was calculated based on a Hertz-based fit using a built-in code written in Igor 6.34A software.

Animal studies. Female NOD/SCID mice (8–9 weeks old) were purchased from Jackson Laboratory, and 8–9-week-old female nude mice were provided by the animal care program at University of California, San Diego. The mice were housed in a special pathogen-free room under standard 12:12-h light:dark cycle, fed with standard rodent chow and water ad libitum, and randomized before experiments. The sample size choice was not pre-determined for each experiment. When comparing RAP2-MST1/2-KO MCF10A cells and MST1/2-dKO MCF10A cells, a total of 5×10^6 MCF10A cells in 50% high concentration Matrigel (BD Bioscience) dissolved in PBS were subcutaneously inoculated into a NOD/SCID mouse. When comparing RAP2-KO MCF10A and MCF10A-T cells with wild-type MCF10A and MCF10A-T cells, 5×10^6 cells in 50% Matrigel were injected into nude mice subcutaneously. When comparing RAP2-KO with wild-type MCF7 cells, 2×10^6 MCF7 cells and 4×10^5 LOX-expressing or control NIH3T3 cells in 50% Matrigel/PBS were co-injected subcutaneously into nude mice. For the semi-synthetic hyaluronan-derived hydrogels, 2×10^6 MCF7 cells in 20 µl PBS suspension were embedded into 200 µl of soft and stiff formulations as described in the preparation of hydrogels. After brief gelation (5 min) at room temperature, the cell-laden hydrogels were subcutaneously injected into nude mice. The investigators were blinded to group allocations during data collection and analyses. All procedures followed the NIH guidelines for the care and use of laboratory animals and the IACUC at the University of California, San Diego approved the experiments. For subcutaneous tumour growth, the maximum single tumour cannot exceed 2 cm in diameter in mice according to the guidelines provided by the animal care program at University of California, San Diego, and no experiments in this study generated tumour burden over this limit.

RNA interference. The lentiviral vectors pLKO.1-hygromycin (Addgene 24150) and pLKO.1-Blasticidin (Addgene 26655) were used to clone the following sense sequences to knock down human YAP or TAZ: negative control: CCTA AGGTAAAGTCGCCCTCG (cloned into pLKO.1-hygromycin and pLKO.1-blasticidin); YAP#1: GCCACCAAGCTAGATAAAGAA (pLKO.1-hygromycin); YAP#2: GACATCTTCTGGTCAGAGATA (pLKO.1-hygromycin); TAZ#1: GCGTTCTTGTGACAGATTATA (pLKO.1-blasticidin); TAZ#2: GCTCATGA GTATGCCCAATGCG (pLKO.1-blasticidin). The resulting plasmids were used to package lentiviruses and the target cells were infected with an MOI of 0.5.

Duplex siRNAs targeting PLCγ1 and PLD1/2 were purchased from Integrated DNA Technologies, Inc. and transfected into cells with RNAiMAX (ThermoFisher Scientific). The sequences are as follows:

PLCγ1 #1 sense strand 5'-rGrArCrUrCrArUrCrArGrCrUrArCrUrArUrGrArGrArAAC-3'; anti-sense strand 5'-rGrUrUrUrCrUrCrArUrArGrUrArGrCrUrArUrGrArGrUrCrArA-3'.

PLCγ1 #2 sense strand 5'-rGrGrCrArArGrArGrUrUrCrUrCrUrCrArGrUrArCrArATC-3'; anti-sense strand 5'-rGrArUrUrGrUrArUrGrArArGrGrArArCrUrCrUrCrUrUrGrCrUrU-3'. PLD1 #1 sense strand 5'-rGrUrGrGrUrArArArUrUrArCrUrUrCrUGT-3'; antisense strand 5'-rArCrArGrArArUrGrArUrArUrUrUrUrCrCrArCrUrG-3'.

PLD1 #2 sense strand 5'-rArCrUrGrGrArArGrArUrUrArCrUrUrGrArArArGrATA-3'; anti-sense strand 5'-rUrArUrCrUrUrUrGrUrCrArGrUrArUrCrUrUrCrUrGrArGrUrG-3'.

PLD2 #1 sense strand 5'-rArArCrCrArArGrArGrArArUrArCrCrGrUrCrArUrUTT-3'; anti-sense strand 5'-rArArArUrGrArCrGrGrUrArUrUrCrUrUrCrUrUrGrGrUrUrGrU-3'.

PLD2 #2 sense strand 5'-rCrUrCrUrArCrArUrUrGrArGrArUrCrArGrUrCrUrUUA-3'; anti-sense strand 5'-rUrGrArArGrArArCrUrGrArUrUrCrUrCrArUrGrUrArGrArGrA-3'.

Quantitative real-time PCR. Total RNAs were extracted with a kit from Qiagen. Reverse transcription was performed with iScript from Bio-Rad. Real-time PCR was performed with the Applied Biosystems 7300 with primers targeting *CTGF*,

CYR61, and *ANKRD1*: *CTGF*-Forward: 5'-CCAATGACAACGCCTCCTG-3', *CTGF*-Reverse: 5'-TGGTGCAGCCAGAAAGCTC-3'; *CYR61*-Forward: 5'-AGCCTCGCATCCTATACAACC-3', *CYR61*-Reverse: 5'-TTCTTTCACAA GCGGCACTC-3'; *ANKRD1*-Forward: 5'-GTGTAGCACCAGATCCATCG-3', *ANKRD1*-Reverse: 5'-CGGTGAGACTGAACCGCTAT-3'. The gene expression was normalized to *GAPDH*: Forward: 5'-TGCACCACCAACTGCTTAGC-3'; Reverse: 5'-GGCATGGACTGTGGTCATGAG-3'.

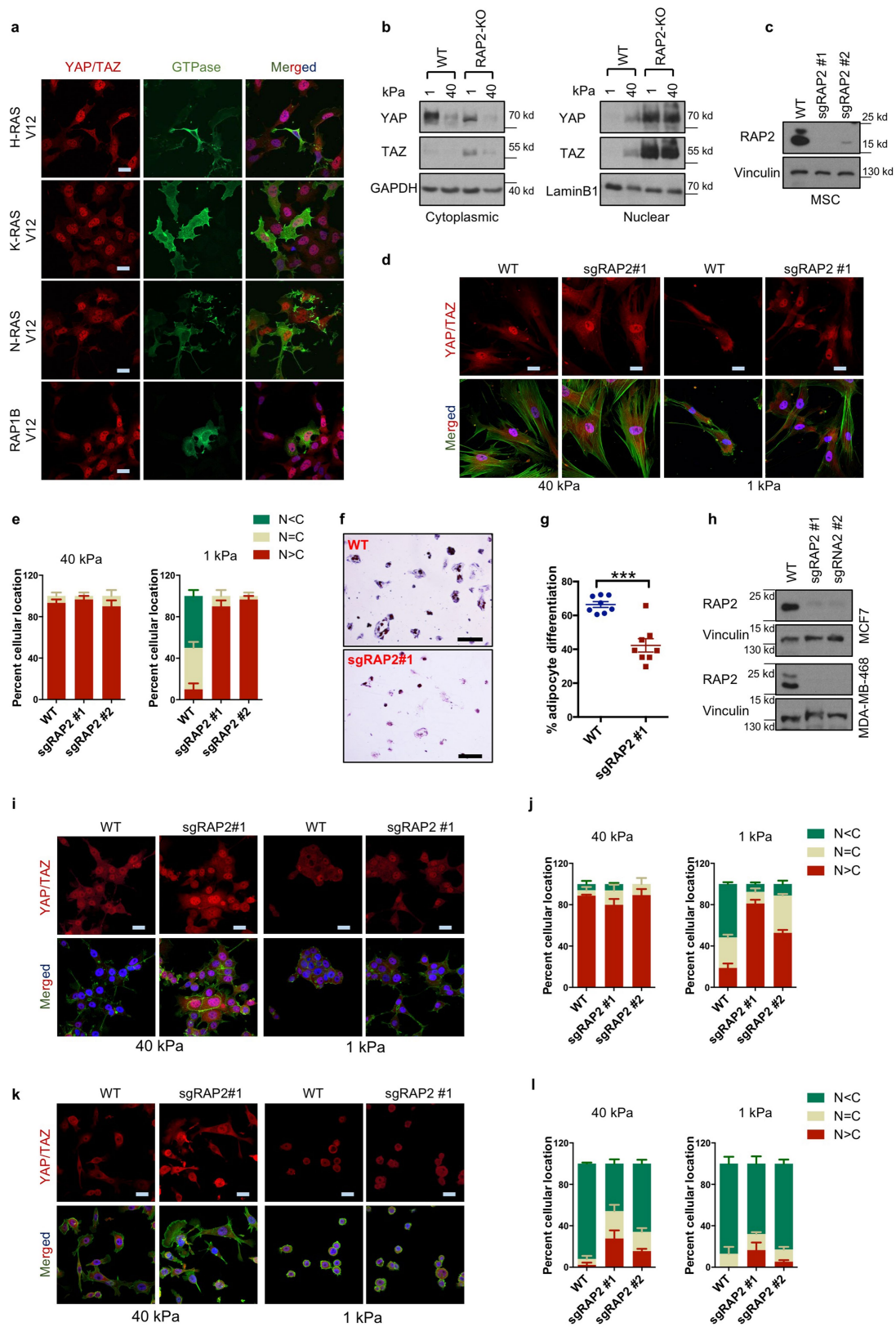
RNA sequencing and bioinformatics analysis. Total RNAs were extracted using TRIzol (Thermo Fisher Scientific) from HEK293A cells seeded on high and low stiffness fibronectin-coated hydrogels. Three replicates for each sample were generated and analysed. The resulting RNA was then used to prepare libraries using Illumina TruSeq Stranded mRNA Library Prep Kit Set A (Illumina, RS-122-2101) or Set B (Illumina, RS-122-2102). The libraries were sequenced using Illumina HiSeq 4000 (single-end 50-bp reads). Reads were aligned to the hg19 reference genome using STAR³². Only uniquely mapped reads were kept for further analysis. Number of reads for each gene were counted using htseq-count³³ according to Gencode human annotation release 24. DeSeq2³⁴ was used to identify differential expressed genes with default parameters. Genes with adjusted *P* value <0.1 were considered to be differentially expressed. GO and KEGG enrichment analysis of differential expressed genes was performed using DAVID³⁵.

Statistical analysis. Microsoft Excel was used for *t*-tests, and Graphpad Prism v6 was used for two-way ANOVA tests. When *P* is smaller than 0.0001, Graphpad

Prism v6 does not provide a precise *P* value and instead only shows a range of *P* < 0.0001.

Data availability. Source Data for Figs. 1, 2, 4 and Extended Data Figs. 1–3, 5, 7–9 can be found in the online version of the paper. For uncropped images of western blot data, see Supplementary Fig. 1. The RNA sequencing data are available in GEO Data Sets with the accession number GSE98547. All other data that support the findings of this study are available upon request from the corresponding author.

29. Wen, J. H. et al. Interplay of matrix stiffness and protein tethering in stem cell differentiation. *Nat. Mater.* **13**, 979–987 (2014).
30. Tse, J. R. & Engler, A. J. Preparation of hydrogel substrates with tunable mechanical properties. *Curr. Protoc. Cell Biol.* Ch. 10, Unit 10.16 (2010).
31. Kaushik, G., Fuhrmann, A., Cammarato, A. & Engler, A. J. In situ mechanical analysis of myofibrillar perturbation and aging on soft, bilayered *Drosophila* myocardium. *Biophys. J.* **101**, 2629–2637 (2011).
32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
33. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
35. Dennis, G. Jr et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, 3 (2003).

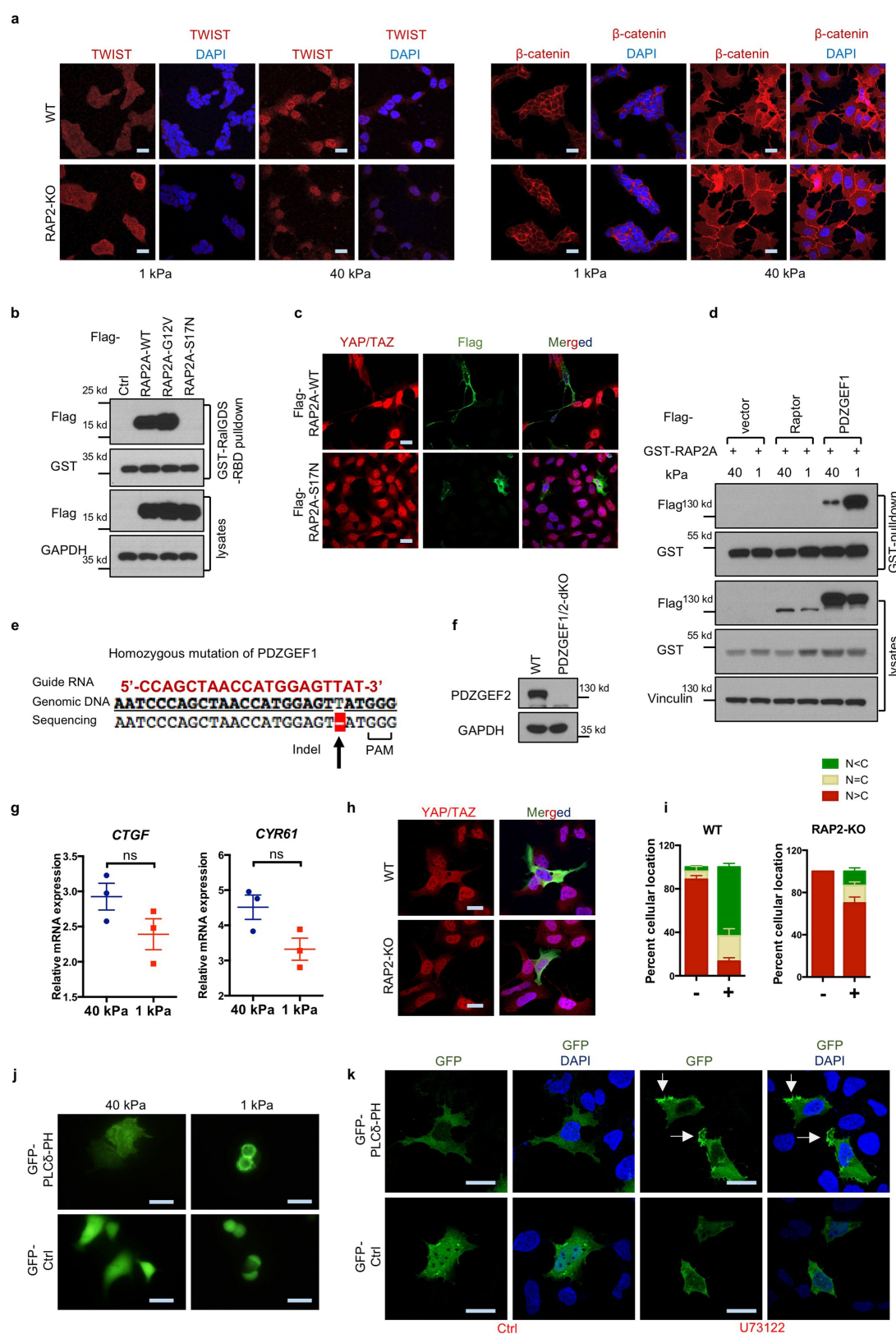


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | RAP2 is involved in regulation of YAP and TAZ by matrix stiffness.

a, Localization of YAP and TAZ is not significantly affected by overexpression of H-RAS, K-RAS, N-RAS, or RAP1B. HEK293A cells were cultured on high-stiffness hydrogels. Overexpression of HA-tagged H-RAS, K-RAS, N-RAS, and Flag-tagged RAP1B are indicated. Merged, combined signals of YAP and TAZ (red), transfected small GTPases (green), and DAPI (blue, staining for DNA). Scale bars, 25 μm . The images are representative of two independent experiments with similar results. **b**, Subcellular fractionation of wild-type and RAP2-KO HEK293A cells at low or high stiffness. The images are representative of two independent experiments with similar results. **c**, CRISPR-mediated deletion of RAP2A, RAP2B and RAP2C in adipocyte-derived MSCs by lentiviral transduction. sgRNAs targeting RAP2A, RAP2B, and RAP2C (sgRAP2) were individually cloned into lentiCRISPR v2 (Addgene #52961) plasmids. sgRAP2 #1 and #2 were two sets of three sgRNAs targeting RAP2A, RAP2B, and RAP2C with unique sequences for each RAP2 isoform. Adipocyte-derived MSCs were infected with the sgRAP2 lentiviruses with an MOI of 10 and selected by puromycin. After puromycin selection, the pooled cells were examined for RAP2 protein expression. The images are representative of two independent experiments with similar results. **d**, RAP2 is required for the low stiffness-induced cytoplasmic localization of YAP and TAZ in MSCs. MSCs with CRISPR-mediated deletion of RAP2A, RAP2B and RAP2C were seeded onto 40 kPa and 1 kPa collagen-coated hydrogels, cultured for 24 h, and then stained for YAP and TAZ. The results are representative of three biologically independent samples showing similar results. Merged, combined signals of YAP and TAZ (red), F-actin (green), and DAPI (blue).

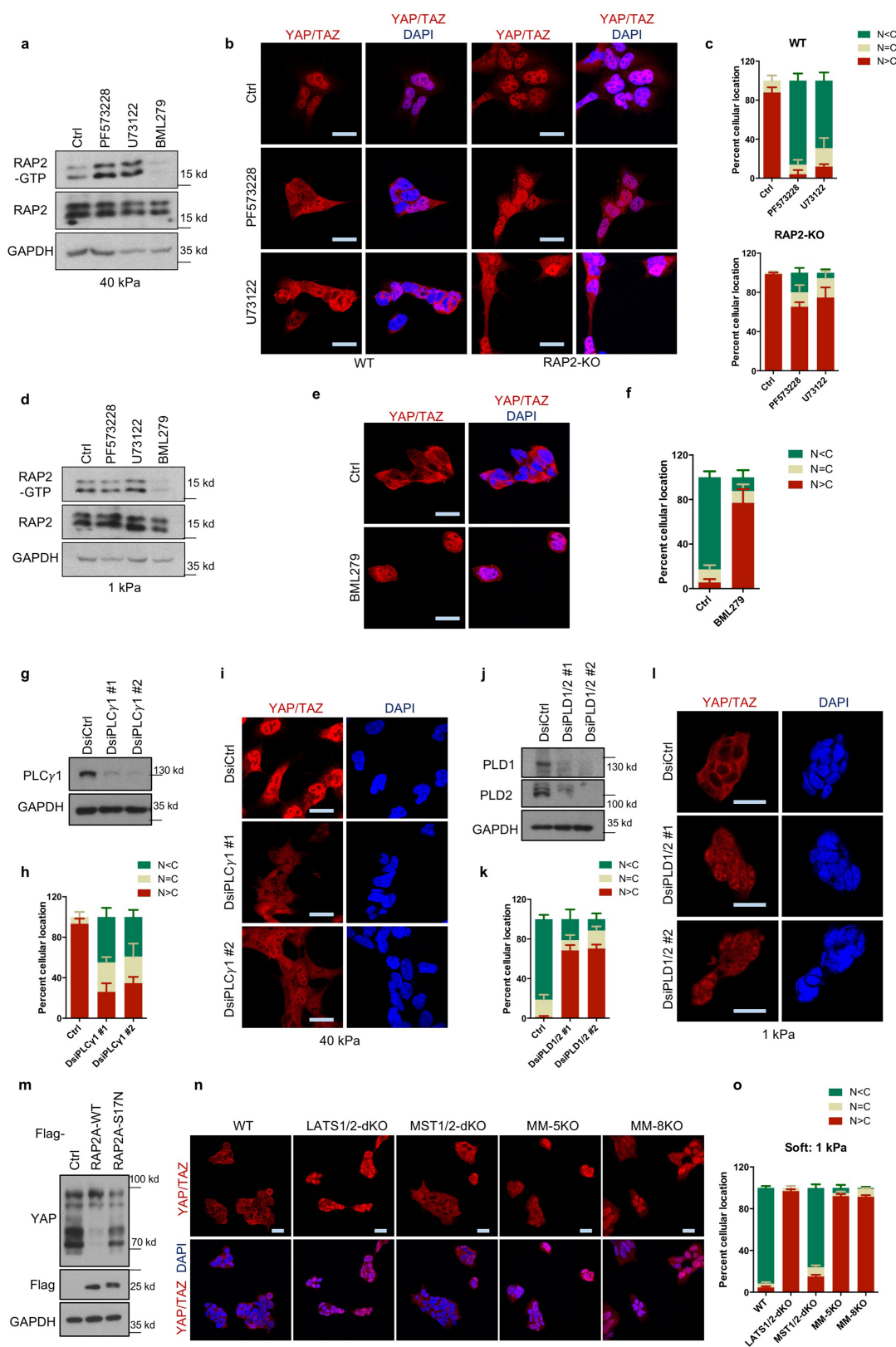
Scale bars, 25 μm . **e**, Distribution of YAP and TAZ localization presented as mean + s.e.m. for cells with more nuclear ($N > C$), more cytoplasmic ($N < C$), or even ($N = C$) distribution of YAP and TAZ. $n = 3$ biologically independent samples. Scale bars, 25 μm . **f**, RAP2 has an important role in adipocyte differentiation at low stiffness. Representative images of Oil-Red O staining of adipocyte-derived MSCs that were treated with adipocyte differentiation medium for 15 days. The MSCs were grown on 1 kPa hydrogels. Scale bars, 200 μm . **g**, Quantification of Oil-Red O-positive cells. The results are presented as mean \pm s.e.m. Two-tailed t -test, $n = 8$ biological independent samples, *** $P = 0.00026$. **h**, CRISPR-mediated deletion of RAP2A, RAP2B and RAP2C in MCF7 and MDA-MB-468 by lentiviral transduction. The experiments were performed similarly to those in **c**. The images are representative of two independent experiments with similar results. **i**, RAP2 deletion promotes nuclear localization of YAP and TAZ in MCF7 cells at low stiffness. The images are from three independent experiments showing similar results. Merged, combined signals of YAP and TAZ (red), F-actin (green), and DAPI (blue). Scale bars, 25 μm . **j**, Quantification of YAP and TAZ localization in MCF7 cells. The distribution of YAP and TAZ localization is presented as mean + s.e.m. $n = 3$ biologically independent samples. **k**, YAP and TAZ are not significantly regulated by matrix stiffness in MDA-MB-468 cells. The images are from three independent experiments showing similar results. Merged, combined signals of YAP and TAZ (red), F-actin (green), and DAPI (blue). Scale bar, 25 μm . **l**, Quantification of YAP and TAZ localization in MDA-MB-468 cells. The distribution of YAP and TAZ localization is presented as mean + s.e.m. $n = 3$ biologically independent samples.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Activation of RAP2 at low matrix stiffness involves PDZGEF1 and PDZGEF2. **a**, RAP2 has no effect on localization of TWIST1 and β -catenin at different matrix rigidities. Immunofluorescence staining of TWIST1 and β -catenin in wild-type and RAP2-KO HEK293A cells grown on soft (1 kPa) and stiff (40 kPa) fibronectin-coated hydrogels. Low stiffness induced cytoplasmic localization of TWIST1 similarly in wild-type and RAP2-KO cells. Stiffness had no significant effect on β -catenin localization in HEK293A cells. Scale bars, 25 μ m. The images are representative of three independent experiments with similar results. **b**, GST-RalGDS-RBD specifically binds to the active RAP2A in the pull-down assay. HEK293A cells were transfected with plasmids expressing wild-type RAP2A, RAP2A-G12V (constitutively GTP-binding), or RAP2A-S17N (GTP-binding deficient) and then seeded onto soft hydrogels. The cells were lysed 24 h after seeding and the lysates were incubated with glutathione agarose beads that were pre-loaded with GST-RalGDS-RBD. The beads were washed and subjected to western blot analyses with the indicated antibodies. The images are representative of two independent experiments with similar results. **c**, GTP-binding and activity of RAP2A is required to induce cytoplasmic translocation of YAP and TAZ. Merged, combined signals from YAP and TAZ (red), Flag (green), and DAPI (blue). Scale bar, 25 μ m. The images are representative of two independent experiments with similar results. **d**, ECM stiffness regulates the interaction between RAP2A and PDZGEF1. GST-RAP2A plasmids were co-transfected with Flag-Raptor (a negative control) or Flag-PDZGEF1 into HEK293A cells. The cells were thereafter seeded on stiff and soft hydrogels. Twenty-four hours after seeding, the cells were lysed and the lysates were incubated with glutathione agarose beads for 6 h. Then the beads were washed and subjected to western blot analyses. The images are representative of two independent experiments with similar results. **e**, Sanger DNA sequencing

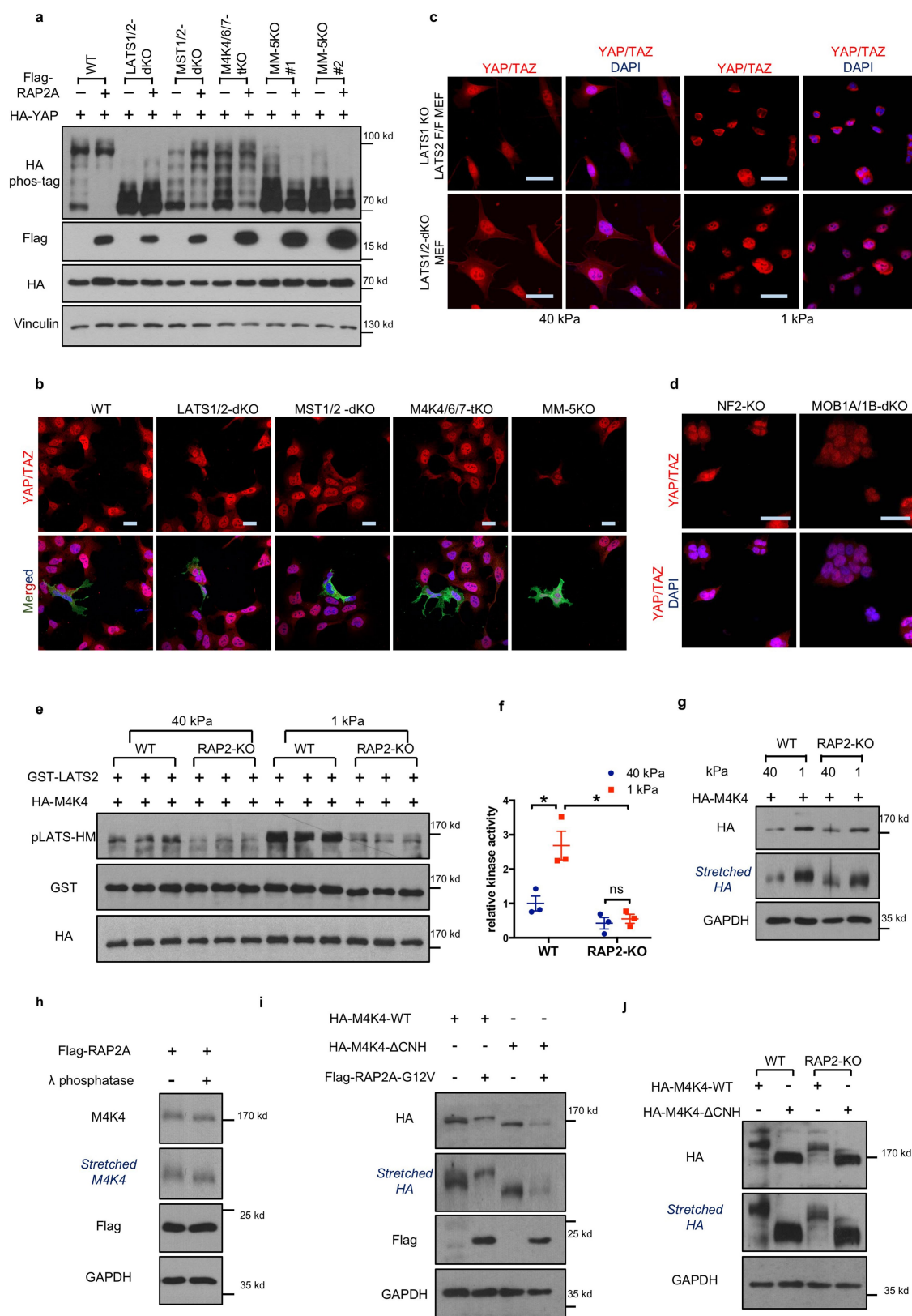
confirmed the homozygous deletion of a 'T' nucleotide in PDZGEF1 genomic DNAs in PDZGEF1/2-dKO HEK293A cells. **f**, Western blot showing the absence of PDZGEF2 expression in PDZGEF1/2-dKO HEK293A cells. The images are representative of two independent experiments with similar results. **g**, The repression of the YAP and TAZ target genes *CTGF* and *CYR61* by low stiffness was compromised in PDZGEF1/2-dKO cells. Expression of *CTGF* and *CYR61* in PDZGEF1/2-dKO HEK293A cells on soft and stiff matrices was measured by qPCR. ns, not significant, two-tailed *t*-test, $n = 3$ biologically independent samples. Data are represented as mean \pm s.e.m. **h**, RAP2 is required for PDZGEF1 to induce cytoplasmic localization of YAP and TAZ. Immunofluorescence showing localization of YAP and TAZ after ectopic expression of Flag-tagged PDZGEF1 in wild-type and RAP2-KO cells at high stiffness. Merged, combined signals from Flag-PDZGEF1 (green), YAP and TAZ (red), and DAPI (blue). Scale bars, 25 μ m. **i**, Quantification of the results in **h**. + denotes Flag-PDZGEF1-transfected cells, - denotes cells that were not transfected. Data are represented as mean + s.e.m. $n = 3$ biologically independent samples. **j**, Stiffness influences cellular PtdIns(4,5) P_2 . The PtdIns(4,5) P_2 reporter GFP-PLC δ -PH domain, which binds to PtdIns(4,5) P_2 , was imaged with a Nikon inverted microscope in cells at low or high stiffness. Cells grown at high stiffness display diffuse PtdIns(4,5) P_2 localization whereas cells at low stiffness show enrichment of PtdIns(4,5) P_2 at the plasma membrane. The image is representative of two independent experiments with similar results. Scale bars, 25 μ m. **k**, Inhibition of PLC alters cellular PtdIns(4,5) P_2 distribution. Immunofluorescence of cells treated with 5 μ M PLC inhibitor U73122 at high stiffness. GFP was detected with anti-GFP immunofluorescence. PtdIns(4,5) P_2 enrichment, indicated by arrows, was observed. Scale bars, 25 μ m. The image is representative of two independent experiments with similar results.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | FAK, PLC γ 1, and PLD are involved in RAP2 activation and Hippo pathway activation in response to stiffness. **a**, The FAK inhibitor PF573228 and PLC inhibitor U73122 promote RAP2-GTP loading at high stiffness. The images are representative of two biologically independent experiments with similar results. **b**, RAP2 functions downstream of FAK and PLC to regulate localization of YAP and TAZ. YAP and TAZ were imaged in cells cultured at high stiffness and treated with 10 μ M FAK inhibitor PF573228 or 5 μ M PLC inhibitor U73122. Scale bars, 25 μ m. The image is representative of three (wild-type treated with PF573228 or U73122; RAP2-KO control (Ctrl)) or four (wild-type control; RAP2-KO treated with PF573228 or U73122) biologically independent samples with similar results. **c**, Distribution of localization from **b** as mean + s.e.m. $n = 3$ (wild-type treated with PF573228 or U73122; RAP2-KO control) or four (wild-type control; RAP2-KO treated with PF573228 or U73122) biologically independent samples. **d**, The PLD1 and PLD2 inhibitor BML279 suppresses RAP2-GTP binding at low stiffness. The images are representative of two independent experiments with similar results. **e**, Inhibition of PLD increases nuclear YAP and TAZ in cells at low stiffness. Cells growing at low stiffness were treated with 5 μ M PLD inhibitor BML279. Scale bars, 25 μ m. Images are representative of seven (control) or five (BML279) biologically independent samples with similar results. **f**, Distribution of localization from **e** presented as mean + s.e.m. $n = 7$ (control) or 5 (BML279) biologically independent samples. **g**, Western blot showing PLC γ 1 knockdown by duplex siRNAs (DsiRNAs). Two independent siRNAs were used. The image is representative of two independent experiments with similar results. **h**, PLC γ 1 knockdown decreases nuclear YAP and TAZ at high stiffness. Quantification of YAP and TAZ localization in PLC γ 1 knockdown and control cells growing on 40 kPa hydrogels. The distribution of YAP and TAZ localization is

presented as mean + s.e.m. $n = 3$ biologically independent samples. **i**, Representative images of YAP and TAZ localization in cells with PLC γ 1 knockdown. Scale bars, 25 μ m. The images are representative of three biologically independent experiments with similar results. **j**, Western blot showing knockdown of PLD1 and PLD2. Two independent DsiRNAs were used. The image is representative of two independent experiments with similar results. **k**, Knockdown of PLD1 and PLD2 increases nuclear YAP and TAZ at low stiffness. Quantification of YAP and TAZ localization in PLD1/2 knockdown and control cells growing on 1 kPa hydrogels. The distribution of YAP and TAZ localization is presented as mean + s.e.m. The results are from three biologically independent samples. **l**, Representative images of YAP and TAZ localization in PLD1/2 knockdown cells. Scale bars, 25 μ m. Images are representative of three biologically independent experiments with similar results. **m**, Wild-type, but not GTP-binding-deficient (S17N), RAP2A induces YAP phosphorylation. YAP phosphorylation was analysed by phos-tag SDS-PAGE in HEK293A cells stably expressing wild-type or S17N mutant RAP2A at high stiffness. Images are representative of two independent experiments with similar results. **n**, Hippo pathway components are involved in regulation of YAP and TAZ by stiffness. HEK293A cell lines in which Hippo components were deleted were cultured at low stiffness (1 kPa). Deletion of Hippo components includes LATS1/2-dKO, MST1/2-dKO, MM-5KO (MST1/2-MAP4K4/6/7-5KO), and MM-8KO (MST1/2-MAP4K1/2/3/4/6/7-8KO). Images are representative of three biologically independent experiments with similar results. Scale bars, 25 μ m. **o**, Quantification of immunofluorescence of the samples in **n**. The distribution of YAP and TAZ localization is presented as mean + s.e.m. $n = 3$ biologically independent samples.

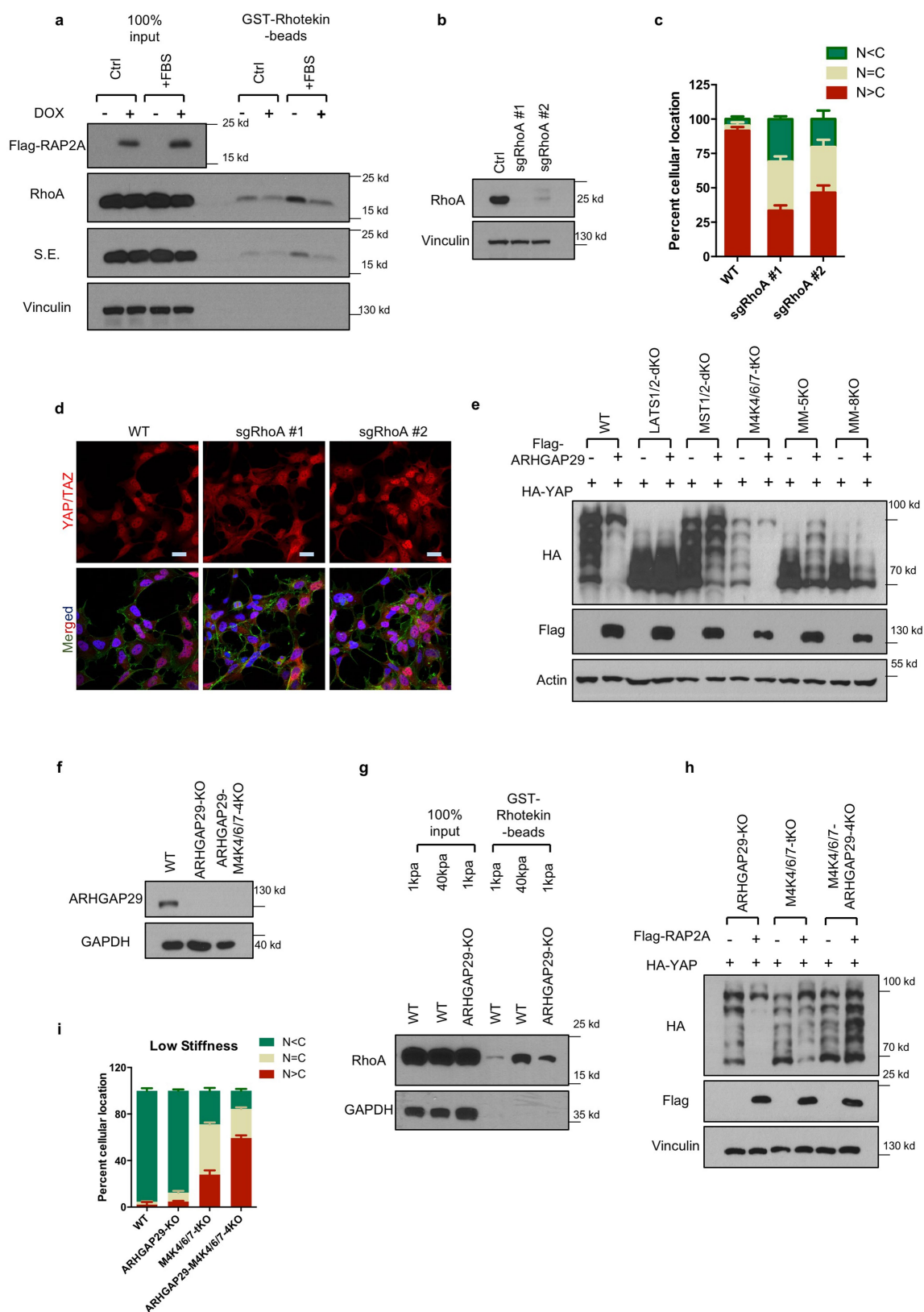


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | RAP2 activates MAP4K4 and induces phosphorylation of YAP.

a, Hippo pathway components are required for RAP2A to induce phosphorylation of YAP and TAZ. RAP2A is overexpressed in the indicated knockout HEK293A cells. Images are representative of two independent experiments with similar results. Two clones of MM-5KO cells were used for this experiment. **b**, RAP2A acts through the Hippo pathway to induce cytoplasmic localization of YAP and TAZ. Flag-RAP2A was transfected into HEK293A cell lines with deletion of different core Hippo pathway components as indicated. Localization of YAP and TAZ was analysed by immunofluorescence. Merged, combined signals from Flag-RAP2A (green), YAP and TAZ (red), and DAPI (blue). Scale bars, 25 μm . Images are representative of three biologically independent experiments with similar results. **c**, Localization of YAP and TAZ in *LATS1*^{-/-}*LATS2*^{flax/flax} (*LATS1*-KO *LATS2*-F/F) and *LATS1*^{-/-}*LATS2*^{-/-} (*LATS1/2*-dKO) mouse embryonic fibroblasts (MEFs). Scale bars, 25 μm . Images are representative of two biologically independent experiments with similar results. **d**, Localization of YAP and TAZ in NF2-KO and MOB1A/1B-dKO HEK293A cells. Scale bars, 25 μm . Images are representative of two biologically independent experiments with similar results. **e**, Low matrix stiffness activates MAP4K4 in a RAP2-dependent manner. Plasmids expressing HA-tagged MAP4K4 were transfected into wild-type and RAP2-KO HEK293A cells. Twenty-four hours after seeding on 40 kPa or 1 kPa hydrogels, HA-MAP4K4 proteins were immunoprecipitated and then used for an in vitro kinase assay, in which recombinant full-length GST-tagged LATS2 protein was used as a substrate. Phosphorylation of LATS2 by MAP4K4 was detected with a

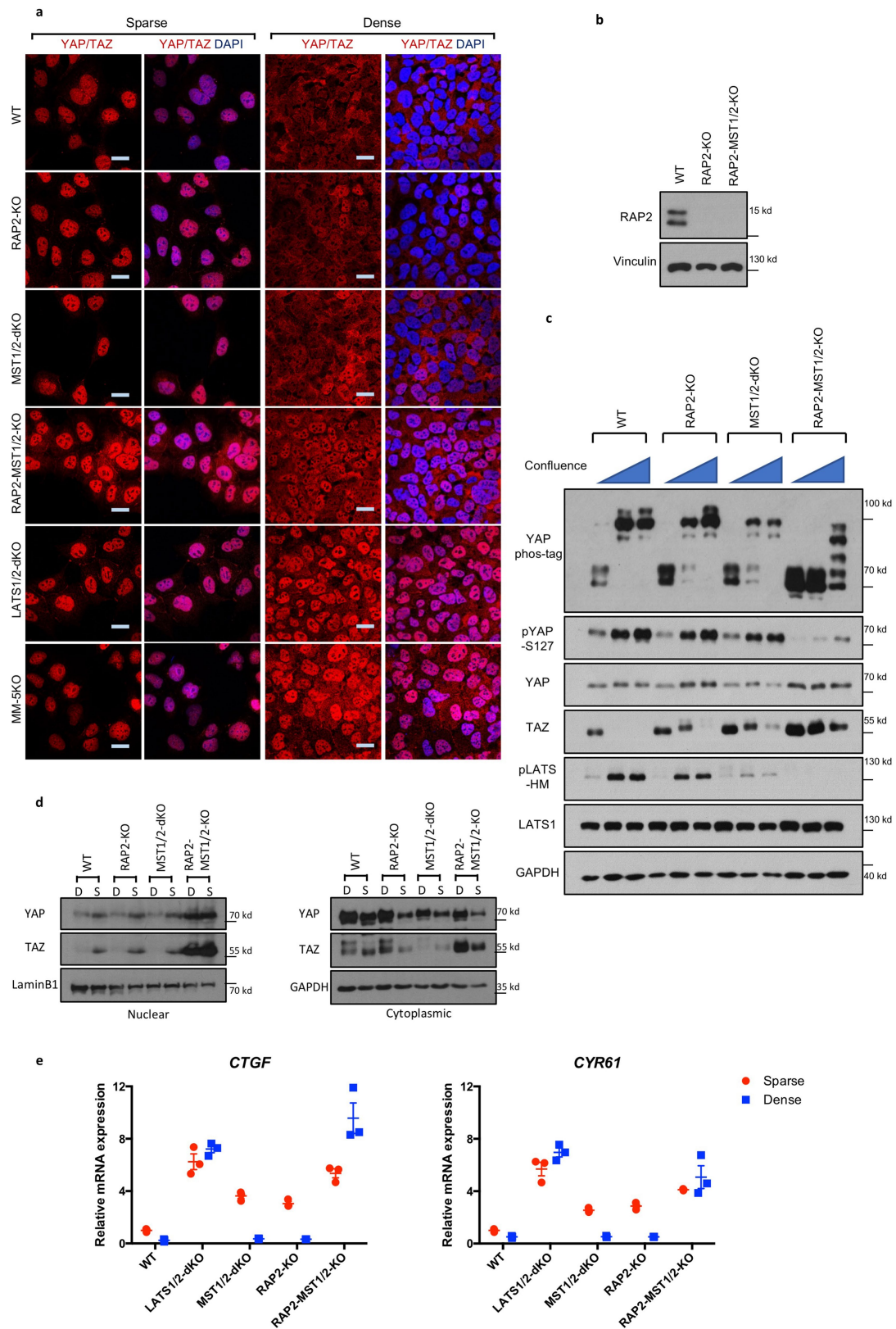
phosphospecific antibody recognizing the phosphorylated hydrophobic motif of LATS1 and LATS2. Results are from three biologically independent experiments. **f**, Quantification of the kinase assay (**e**) shown as mean \pm s.e.m. Relative LATS2 phosphorylation was normalized to protein levels of HA-MAP4K4 and defined as MAP4K4 kinase activity. $n = 3$ biologically independent samples, two-tailed *t*-test, $*P = 0.027$ (wild-type, 1 kPa versus RAP2 KO, 1 kPa) or 0.037 (wild-type, 1 kPa versus wild-type, 40 kPa); ns, not significant. **g**, The lysates of wild-type and RAP2-KO HEK293A cells growing on stiff and soft hydrogels were analysed by western blot for MAP4K4 migration. A 'stretched' image was generated by vertically extending the same western image directly above it in order to better visualize the MAP4K4 mobility shift. The image is representative of two independent experiments with similar results. **h**, Phosphatase treatment increases MAP4K4 migration on SDS-PAGE. Lambda phosphatase was used to treat cell lysates of RAP2A-expressing HEK293A cells before western blotting. These results indicate that the altered migration of MAP4K4 was correlated with its phosphorylation. The image is representative of two independent experiments with similar results. **i**, RAP2A promotes MAP4K4 phosphorylation (slower band migration) dependent on its citron domain. Flag-RAP2A-G12V plasmid was co-transfected with wild-type MAP4K4 or the citron domain deletion mutant (ΔCNH) into HEK293A cells. The image is representative of two independent experiments with similar results. **j**, RAP2 is required for the reduction in MAP4K4 mobility. Wild-type and ΔCNH mutant MAP4K4 were transfected into wild-type and RAP2-KO HEK293A cells. The image is representative of two independent experiments with similar results.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | RAP2 inhibits RhoA GTPase through ARHGAP29. **a**, RAP2A expression inhibits endogenous RhoA GTP-binding. HEK293A cells with doxycycline-inducible expression of RAP2A were established and the expression of Flag-RAP2A was induced by doxycycline. RhoA activity was determined by a GST-Rhotekin-RBD pull-down assay. s.e. denotes short exposure of the RhoA western blot. The image is representative of two independent experiments with similar results. **b**, Western blot confirms CRISPR-mediated RhoA gene editing. HEK293A cells were transfected with CRISPR plasmids targeting RhoA and selected with puromycin for 3 days. Two sgRNAs were used to generate two RhoA knockout pools (sgRhoA #1 and #2)¹⁶. The image is representative of two independent experiments with similar results. **c**, CRISPR-mediated deletion of RhoA leads to increased cytoplasmic localization of YAP and TAZ in HEK293A cells at high stiffness. The localization distribution is presented as mean + s.e.m. $n = 6$ biologically independent samples. **d**, Representative immunofluorescence images from **c**. Scale bars, 25 μm . **e**, ARHGAP29 induces YAP phosphorylation in a Hippo pathway-dependent manner. MM-5KO, a HEK293A clone with deletion of MST1, MST2, MAP4K4, MAP4K6 and MAP4K7. MM-8KO, deletion of MST1, MST2, MAP4K1, MAP4K2, MAP4K3, MAP4K4, MAP4K6 and MAP4K7. YAP phosphorylation was detected

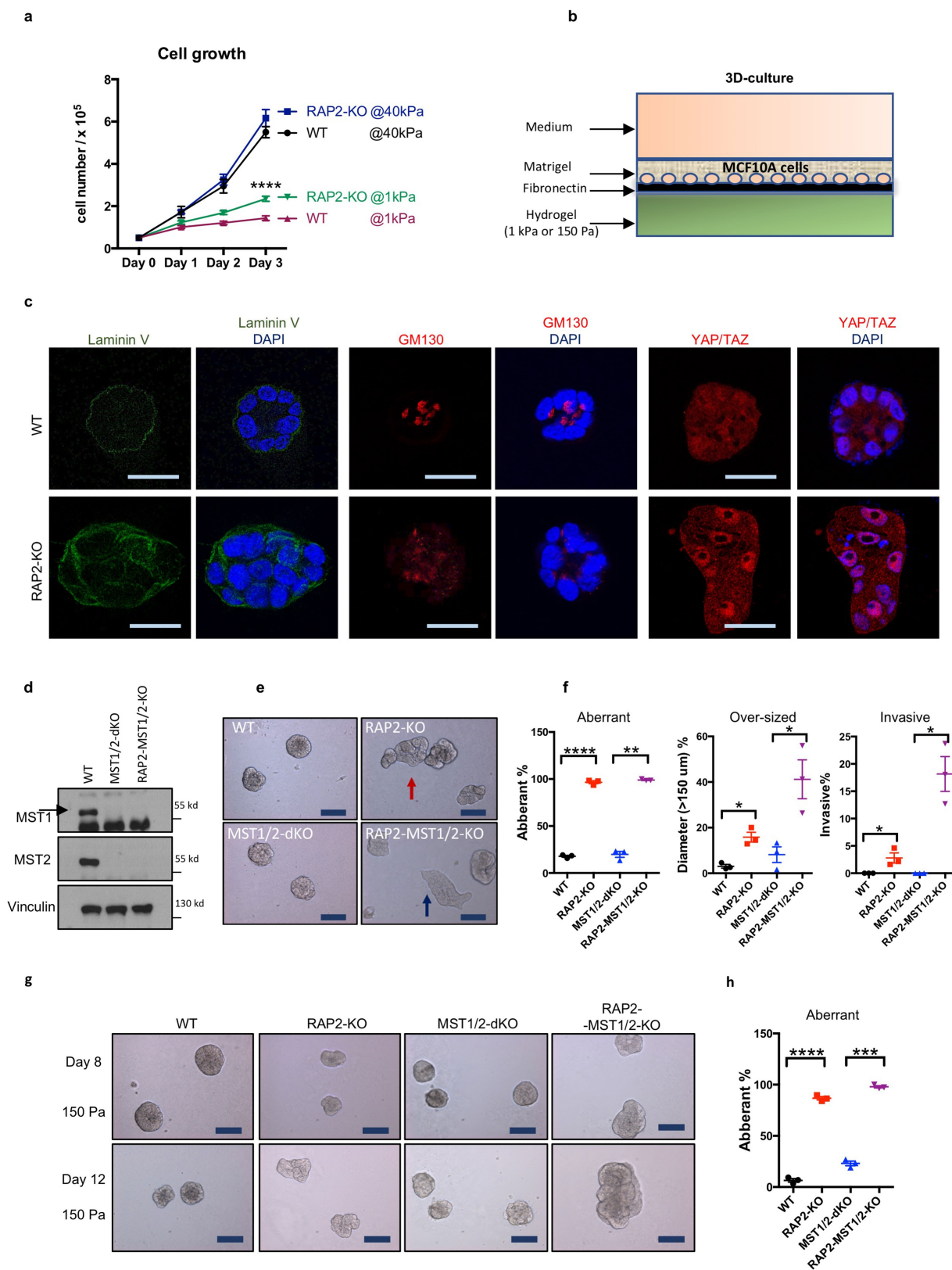
by phos-tag gels. The image is representative of two independent experiments with similar results. **f**, Immunoblot showing deletion of ARHGAP29 in HEK293A wild-type and MAP4K4/6/7-tKO cells. The image is representative of two independent experiments with similar results. **g**, Deletion of ARHGAP29 compromises inhibition of RhoA by low matrix stiffness. Wild-type and ARHGAP29-KO HEK293A cells were cultured at the indicated stiffness and then assayed for RhoA activity with a GST-Rhotekin-RBD binding assay. The image is representative of two independent experiments with similar results. **h**, Combined deletion of ARHGAP29, MAP4K4, MAP4K6 and MAP4K7 abolishes YAP phosphorylation induced by RAP2A. HA-YAP was co-transfected with vector or Flag-RAP2A into HEK293A cells cultured at high stiffness. HA-YAP phosphorylation was detected by phos-tag SDS-PAGE. The image is representative of two independent experiments with similar results. **i**, Combined deletion of ARHGAP29, MAP4K4, MAP4K6 and MAP4K7 blocks low stiffness-induced cytoplasmic localization of YAP. Quantification of YAP and TAZ localization in HEK293A cells with deletion of ARHGAP29 and/or MAP4K4, MAP4K6 and MAP4K7 at low stiffness in Fig. 3e. The YAP and TAZ localization distribution is presented as mean + s.e.m. $n = 3$ biologically independent samples.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | RAP2 contributes to cell density-induced inactivation of YAP and TAZ. **a**, Cytoplasmic translocation of YAP and TAZ caused by cell contact involves RAP2 and the Hippo pathway. Scale bars, 25 μ m. MM-5KO, a HEK293A clone with deletion of MST1, MST2, MAP4K4, MAP4K6 and MAP4K7. RAP2-MST-KO, deletion of RAP2A, RAP2B, RAP2C, MST1 and MST2. The images are representative of three biologically independent experiments with similar results. **b**, Western blot showing absence of RAP2 proteins in RAP2-KO and RAP2-MST1/2-KO cells. Note that part of these results is shown in Fig. 1b and are from the same experiment. The image is representative of two independent experiments with similar results. **c**, Phosphorylation of YAP and TAZ induced by cell contact requires RAP2, MST1 and MST2. The western blot shows phosphorylation of YAP in cells with low, medium, or high

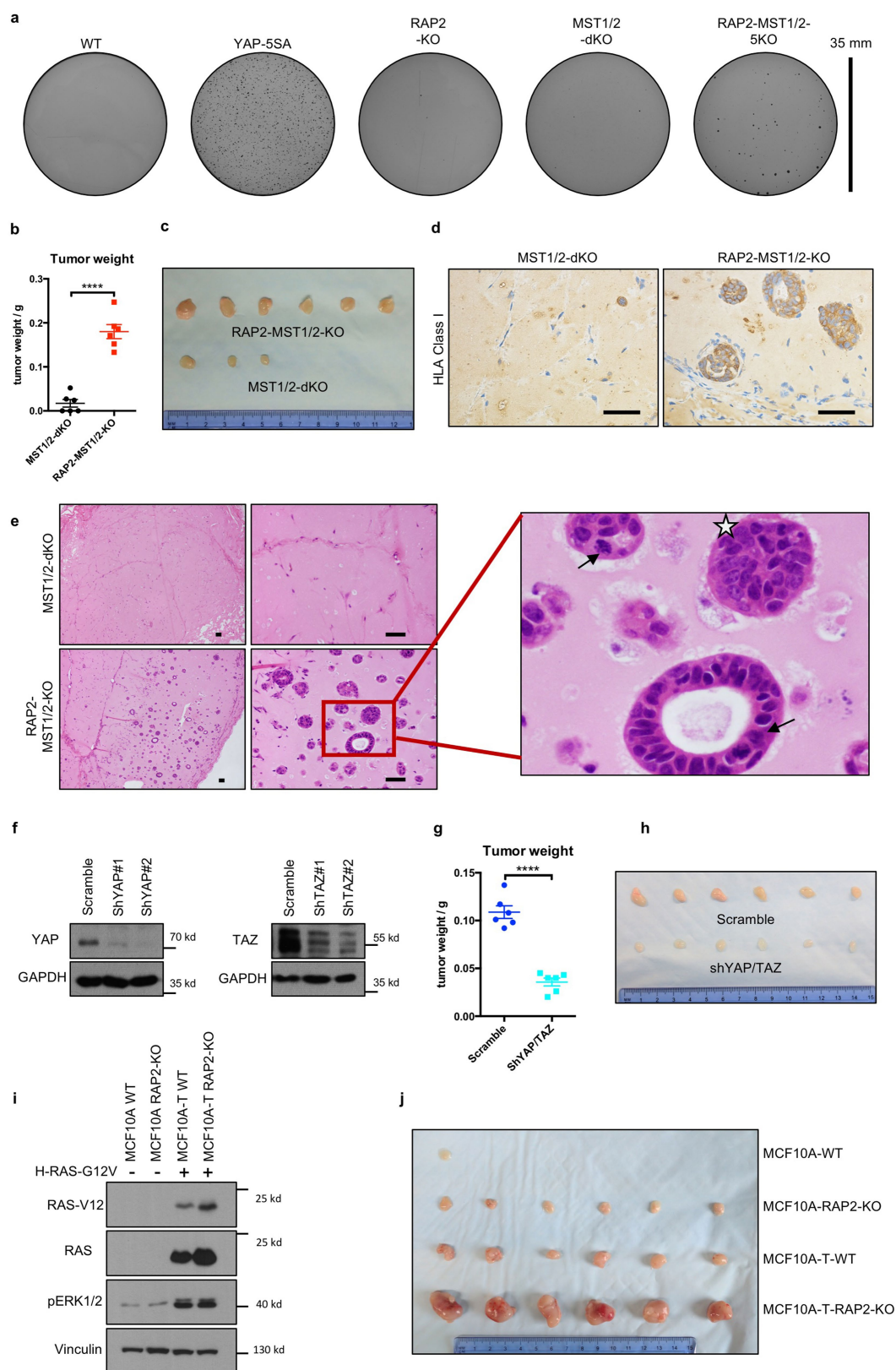
confluence. The image is representative of two independent experiments with similar results. **d**, Deletions of RAP2A, RAP2B, RAP2C, MST1 and MST2 interfere with regulation of YAP and TAZ by cell density. Subcellular fractionations were performed for wild-type, RAP2-KO, MST1/2-dKO and RAP2-MST1/2-KO HEK293A cells cultured at low (S, sparse) or high density (D, dense). GAPDH and LaminB1 are markers for the cytoplasmic and nuclear fraction, respectively. The image is representative of two independent experiments with similar results. **e**, RAP2A, RAP2B, RAP2C, MST1 and MST2 are required for regulation of YAP and TAZ target genes by cell density. qPCR was performed to determine the expression of the YAP and TAZ target genes *CTGF* and *CYR61* in the above cells at low or high confluence. Data are represented as mean \pm s.e.m. $n = 3$ biologically independent samples.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | RAP2 prevents aberrant acinus growth in MCF10A cells on soft matrices. **a**, RAP2 deletion selectively enhances HEK293A cell growth on soft matrices. Wild-type and RAP2-KO HEK293A cells were seeded on stiff or soft matrices, and cell numbers were recorded as mean \pm s.e.m. every day. Two-way ANOVA test, **** $P < 0.0001$, wild-type versus RAP2-KO cells cultured at 1 kPa, $n = 3$ biologically independent samples. **b**, Model of 3D culture of MCF10A cells. **c**, RAP2 deletion causes abnormal acinus growth and cell polarity defects in MCF10A cells. Immunofluorescence staining of acini for cell polarity markers, Laminin V and GM130, and YAP and TAZ in wild-type and RAP2-KO MCF10A cells cultured for 6 days. Scale bars, 25 μm . Images are representative of three independent experiments with similar results. **d**, Western blot showing deletion of MST1 and MST2 in wild-type and RAP2-KO cells. The arrow indicates the specific band for MST1. The image is representative of two independent experiments with similar results. **e**, Representative images showing acinus formation by various genetically engineered MCF10A cells at 1 kPa. The red arrow indicates aberrant acini. The blue arrow indicates invasive cell morphology.

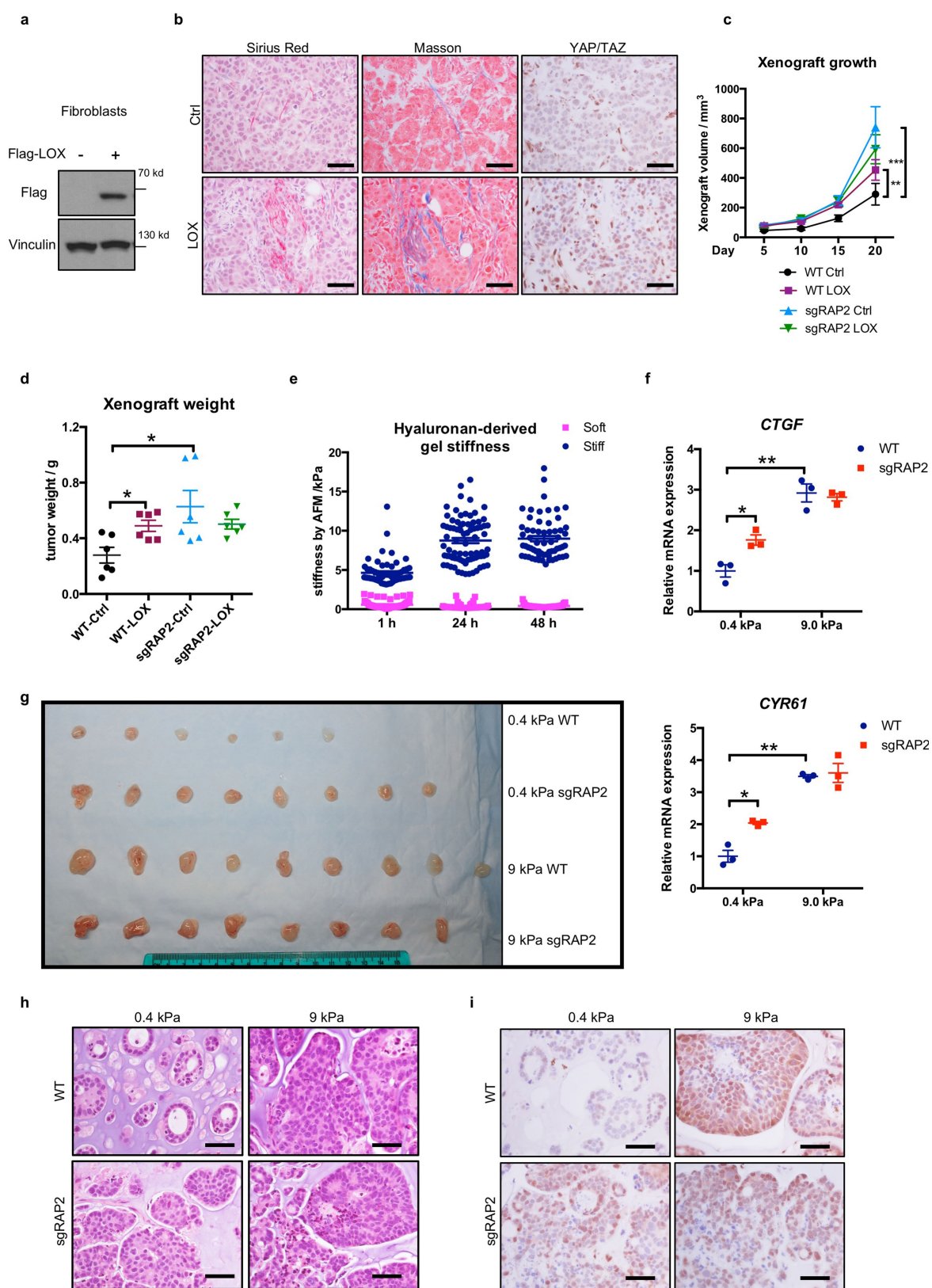
Scale bars, 100 μm . The results are representative of two independent experiments with similar results. **f**, Quantification of aberrant, oversized, and invasive acini in **e**. The percentage of the cells is presented as mean \pm s.e.m. Two-tailed t -tests were used for statistical analyses of aberrant and oversized clones in $n = 3$ biologically independent samples. Aberrant: **** $P = 0.0000062$; *** $P = 0.0012$; oversized: * $P = 0.017$ (wild-type versus RAP2-KO) or 0.046 (MST1/2-dKO versus RAP2-MST1/2-KO). For analysis of invasive clones, one-tailed t -test was used as no invasive clones were observed in either wild-type or MST1/2-dKO cells. * $P = 0.047$ (wild-type versus RAP2-KO) or 0.029 (MST1/2-dKO versus RAP2-MST1/2-KO). $n = 3$ biologically independent samples. **g**, Representative images showing acinus formation by various genetically engineered MCF10A cells at 150 Pa. Scale bars, 100 μm . Images are representative of two biologically independent experiments with similar results. **h**, Quantification of aberrant acini in **g**. The results are from three biologically independent samples. Two-tailed t -tests were used for statistical analyses of aberrant clones in $n = 3$ biologically independent samples. **** $P = 0.000045$; *** $P = 0.00011$.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | RAP2 deletion contributes to aberrant acinus growth and tumorigenesis of MCF10A cells in a YAP- and TAZ-dependent manner. **a**, Representative images showing soft-agar assays of MCF10A cells. Overexpression of the constitutively active YAP(5SA) (mutation of all five LATS1- and LATS2-phosphorylation serines to alanines in YAP) strongly promotes anchorage-independent growth. Combined deletion of RAP2, MST1 and MST2, but not either group alone, also causes anchorage-independent growth of MCF10A cells. Images are representative of three biologically independent experiments with similar results. Scale bar, 35 mm. **b**, RAP2 inhibits tumorigenicity of MST1/2-dKO MCF10A cells. MST1/2-dKO and RAP2-MST1/2-KO MCF10A cells were injected into NOD/SCID mice. Tumour weight on day 32 after injection is presented as mean \pm s.e.m. $n = 6$ biologically independent xenografts, **** $P = 0.000025$, two-tailed t -test. **c**, Representative tumour sizes. Only three very small xenografts were recovered from the initial 6 subcutaneous injections for MST1/2-dKO cells. Images representative of six biologically independent xenografts for each group that were initially made in the NOD/SCID mice. **d**, Immunohistochemistry staining with an antibody recognizing human HLA Class I. Only the acinus structures in the xenografts were formed by MCF10A cells. Stroma cells negative for HLA class I were derived from the host mice. Images are representative of two biologically independent experiments with similar results. Scale bars, 50 μ m. **e**, Haematoxylin and eosin staining of xenografts from MST1/2-dKO MCF10A cells shows that largely hypocellular connective

tissue is observed as stroma from the host animals. By contrast, RAP2-MST1/2-KO xenografts showed MCF10A cell-derived acinar and duct structures exhibiting nuclear polymorphisms, irregular nuclear contour, hyperchromasia, prominent nucleoli (star), and pathological mitosis (arrows). Images representative of two biologically independent experiments with similar results. Scale bars, 50 μ m. **f**, Western blot showing knockdown of YAP or TAZ by lentiviral shRNAs in RAP2-MST1/2-KO MCF10A cells. shYAP#2 and shTAZ#1 were used for the xenograft studies. The image is representative of two independent experiments with similar results. **g**, Knockdown of YAP and TAZ inhibits tumour growth of RAP2-MST1/2-KO MCF10A cells. Tumour weight on day 32 is presented as mean \pm s.e.m. Two-tailed t -test, $n = 6$, **** $P = 0.000010$. **h**, Xenografts from NOD/SCID mice, in which six biologically independent xenografts were generated for each group. **i**, Western blot showing that MCF10A-T cells are generated by expression of the oncogenic mutant H-RAS-G12V. H-RAS-G12V expression activates ERK whereas RAP2 deletion has no effect on ERK. The comparable phosphorylation levels of ERK1 and ERK2 in wild-type and RAP2-KO MCF10A-T cells indicate that the difference in xenograft growth was not due to difference in ERK1 and ERK2 activity. Image representative of two independent experiments with similar results. **j**, MCF10A and MCF10A-T xenografts from nude mice, in which six biologically independent xenografts were generated for each group and yielded similar results.

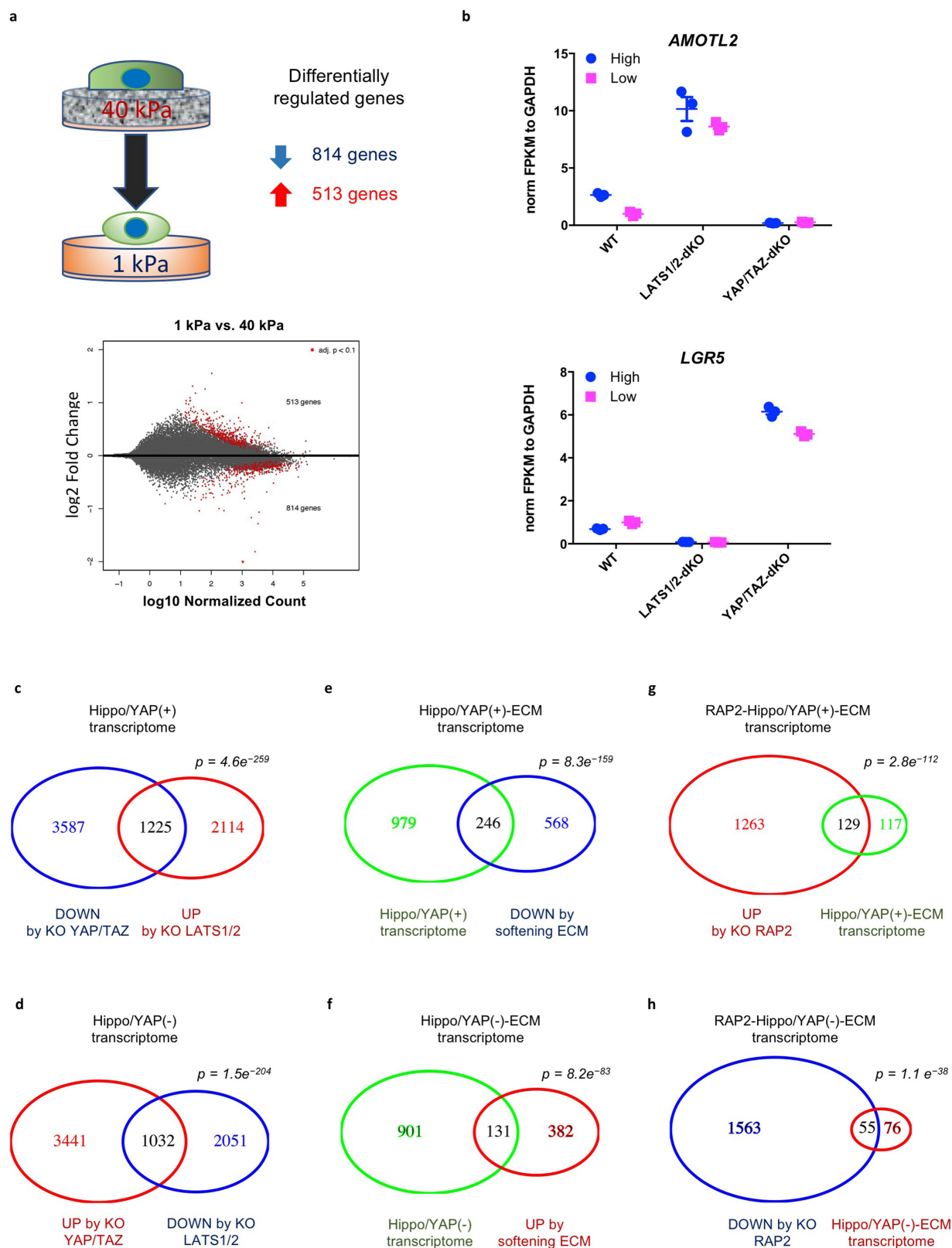


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | RAP2 deletion selectively promotes MCF7 malignancy at low stiffness in vivo.

a, Western blot showing LOX overexpression in NIH3T3 fibroblasts. Ectopic LOX expression promotes cross-linking of ECM proteins and thus increases matrix stiffness²⁸. Image representative of two independent experiments with similar results. **b**, RAP2 is involved in the xenograft growth enhancement caused by LOX-overexpressing fibroblasts. Xenografts were generated by co-injection of 0.4×10^6 fibroblasts (control and Flag-LOX-overexpressing) and 2.0×10^6 MCF7 cells. The xenografts were removed on day 6 and examined for collagen deposition and crosslinking and localization of YAP and TAZ. LOX overexpression led to a woven-like structure of collagen in the xenografts based on Sirius Red staining (red colour for collagen) and Masson staining (blue colour for collagen), and increased nuclear YAP and TAZ. Images representative of three biologically independent experiments with similar results. Scale bars, 50 μm . **c**, LOX-induced tumour growth requires RAP2. Xenografts were generated by co-injection of 0.4×10^6 NIH3T3 cells and 2.0×10^6 MCF7 cells. The growth of the xenografts with different combinations of NIH3T3 and MCF7 cells is shown as mean \pm s.e.m. Deletion of RAP2A, RAP2B and RAP2C promoted MCF7 tumour growth and masked the enhancement induced by co-injected LOX-expressing fibroblasts. **two-way ANOVA test (wild-type MCF7 + wild-type NIH3T3 cells versus wild-type MCF7 + LOX-overexpressing NIH3T3 cells), $n = 6$ biologically independent samples, $P = 0.0027$. ***two-way ANOVA test (sgRAP2 MCF7 + wild-type NIH3T3 cells versus wild-type MCF7 + wild-type NIH3T3 cells), $n = 6$ biologically independent samples, $P = 0.002$. **d**, Tumour weights (mean \pm s.e.m.). *two-tailed t -test, $n = 6$ biologically independent xenografts, $P = 0.014$ (wild-type MCF7 + wild-type NIH3T3 cells versus wild-type MCF7 + LOX-overexpressing NIH3T3 cells) or 0.029 (sgRAP2 MCF7 + wild-type NIH3T3 cells versus wild-type MCF7 + wild-type NIH3T3 cells). **e**, The

elasticity or stiffness of the 'soft' and 'stiff' semi-synthetic hyaluronan-derived gels was measured by atomic force microscopy. Results presented as mean \pm s.e.m. The measurements were made more than 55 times for each stiffness at each time. **f**, RAP2 inhibits expression of the YAP and TAZ target genes *CTGF* and *CYR61* at low stiffness but not at high stiffness. Quantitative real-time PCR analyses of *CTGF* and *CYR61* in wild-type and sgRAP2 MCF7 cells (as in Extended Data Fig. 1g) cultured in vitro for 48 h in soft or stiff hyaluronan gel. Relative mRNA levels presented as mean \pm s.e.m. $n = 3$ biologically independent samples, two-tailed t -test. For *CTGF*, *sgRAP2 versus wild-type at 0.4 kPa, $P = 0.020$; **0.4 kPa versus 9.0 kPa for wild-type, $P = 0.0032$. For *CYR61*, *sgRAP2 versus wild-type, $P = 0.025$; **0.4 kPa versus 9.0 kPa for wild-type, $P = 0.0033$. **g**, Xenograft tumours. For wild-type cells grown at 0.4 kPa, eight independent xenografts were initially generated in nude mice. However, owing to animal deaths, only six xenografts were recovered. For wild-type cells grown at 9 kPa, nine independent xenografts were generated. For sgRAP2 cells grown at both 0.4 kPa and 9 kPa, eight independent xenografts were generated. **h**, RAP2 deletion preferentially promotes MCF7 malignancy at low stiffness. MCF7 xenografts stained with haematoxylin and eosin. Wild-type cells embedded in 0.4 kPa matrix produced mostly tubular and some cribriform structures whereas those embedded in 9 kPa matrix produced mostly solid nests, as well as more marked cellular pleomorphism and nuclear atypia. sgRAP2 cells showed more malignant architecture and morphology at 0.4 ka, while at high stiffness (9 kPa) sgRAP2 and wild-type cells exhibited similar morphology. Images representative of three biologically independent experiments with similar results. Scale bars, 50 μm . **i**, Immunohistochemistry for YAP and TAZ in xenografts. RAP2 deletion increased nuclear YAP and TAZ. Images representative of three biologically independent experiments with similar results. Scale bars, 50 μm .



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | RAP2 mediates regulation of the ECM stiffness transcriptome by the Hippo pathway. **a**, MA plot of wild-type cells in low stiffness versus wild-type cells in high stiffness. Three biologically independent samples were assayed for each condition. Differentially expressed genes (adjusted P value <0.1) are coloured red. P values for differential expression were derived using the Wald test and corrected using the Benjamini–Hochberg procedure with default functions in DESeq2. **b**, Dot plot showing expression of *AMOTL2* and *LGR5*. *AMOTL2* and *LGR5* are YAP- and TAZ-dependent and stiffness-regulated genes. High and low denote high stiffness (40 kPa) and low stiffness (1 kPa). Data are presented as mean \pm s.e.m. $n = 3$ biologically independent samples. **c**, Venn diagram comparing genes downregulated by YAP and

TAZ knockout at high stiffness and genes upregulated by LATS1 and LATS2 knockout at low stiffness. **d**, Venn diagram comparing genes upregulated by YAP and TAZ knockout at high stiffness and genes downregulated by LATS1 and LATS2 knockout at low stiffness. **e**, Venn diagram comparing overlapping genes from **c** and genes downregulated by low stiffness. **f**, Venn diagram comparing overlapping genes from **d** and genes upregulated by low stiffness. **g**, Venn diagram comparing overlapping genes from **e** and genes upregulated by RAP2 knockout at low stiffness. **h**, Venn diagram comparing overlapping genes from **f** and genes downregulated by RAP2 knockout at low stiffness. P values for **c–h**: hypergeometric test. Results represent analyses of three biological replicates for each condition.

SIRT6 deficiency results in developmental retardation in cynomolgus monkeys

WeiQi Zhang^{1,2,3,4,8}, Haifeng Wan^{2,3,4,8}, Guihai Feng^{2,3,4,8}, Jing Qu^{2,3,4,8}, Jiaqiang Wang^{2,3,4}, Yaobin Jing^{1,3}, Ruotong Ren^{1,3,4}, Zunpeng Liu^{2,3}, Linlin Zhang^{2,3}, Zhiguo Chen⁵, Shuyan Wang⁵, Yong Zhao⁶, Zhaoxia Wang⁷, Yun Yuan⁷, Qi Zhou^{2,3,4}, Wei Li^{2,3,4,*}, Guang-Hui Liu^{1,3,4,5,*} & Baoyang Hu^{2,3,4,*}

SIRT6 acts as a longevity protein in rodents^{1,2}. However, its biological function in primates remains largely unknown. Here we generate a SIRT6-null cynomolgus monkey (*Macaca fascicularis*) model using a CRISPR–Cas9-based approach. SIRT6-deficient monkeys die hours after birth and exhibit severe prenatal developmental retardation. SIRT6 loss delays neuronal differentiation by transcriptionally activating the long non-coding RNA *H19* (a developmental repressor), and we were able to recapitulate this process in a human neural progenitor cell differentiation system. SIRT6 deficiency results in histone hyperacetylation at the imprinting control region of *H19*, CTCF recruitment and upregulation of *H19*. Our results suggest that SIRT6 is involved in regulating development in non-human primates, and may provide mechanistic insight into human perinatal lethality syndrome.

To explore the roles of SIRT6 in primates, we generated a biallelic SIRT6-null cynomolgus monkey model in a one-step procedure using a CRISPR–Cas9 system^{3–5} (Extended Data Fig. 1a and Supplementary Table 1). We initially designed six single guide RNAs (sgRNAs) that target conserved domains in the monkey *SIRT6* gene and determined that two sgRNAs, each of which target the deacetylase domain of SIRT6, exhibited a higher targeting efficiency (Extended Data Fig. 1a, b, Supplementary Table 1 and Supplementary Information Guide). Subsequently, the two SIRT6-specific sgRNAs and Cas9 mRNA were injected into 98 monkey zygotes, and 48 embryos that developed and displayed normal morphology were transferred into 12 surrogate mothers. Four surrogates were successfully impregnated, but only three completed the pregnancy cycle (of about 165 days), and each gave birth to one female infant (Fig. 1a, b and Supplementary Table 1). The other surrogate miscarried at the middle of the gestation stage and delivered a nonviable male fetus (Supplementary Table 1).

We evaluated whether the CRISPR–Cas9 system caused mutagenesis in the gene-modified monkeys by performing genomic DNA PCR and Sanger sequencing. Genotyping of 15 different tissues from the three engineered monkeys revealed compound biallelic mutations in *SIRT6* gene, which suggests that highly efficient biallelic mutagenesis occurred in the founder monkeys (Extended Data Fig. 1c–e and Supplementary Table 1). Whole-genome DNA sequencing further confirmed the precise targeting of the *SIRT6* locus in all three monkeys, without detectable off-target effects across the genome (Fig. 1c, Extended Data Fig. 1f–h and Supplementary Table 2). Consistent with disruptions in the *SIRT6* gene, in all the tissues from SIRT6^{−/−} monkeys that we tested—which included the brain, muscle, liver, kidney, lung and skin—the SIRT6 protein was not detected (Fig. 1d, Extended Data Fig. 2a, Supplementary Fig. 1). In addition, the levels of histone 3 acetylated at K56 (H3K56ac)—a major SIRT6 substrate^{1,6,7}—were increased in various tissues of the SIRT6-null monkeys (Fig. 1d and Extended Data

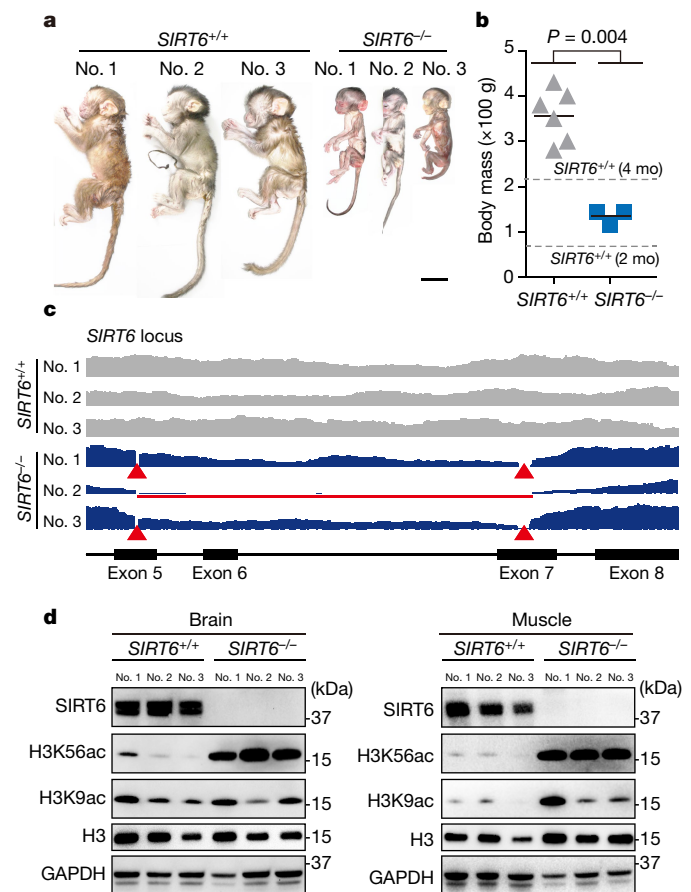


Fig. 1 | Generation and characterization of SIRT6^{−/−} monkeys.

a, Photographs of wild-type monkeys (left) and SIRT6^{−/−} monkeys (right). Scale bar, 5 cm. **b**, Body masses of wild-type and SIRT6^{−/−} monkeys at birth. Grey dashed lines represent body mass values of 2-month-old and 4-month-old wild-type fetuses ($n = 6$ SIRT6^{+/+} monkeys, $n = 3$ SIRT6^{−/−} monkeys). Horizontal lines show the average values for each group; P values were determined by two-sided Student's t -test. **c**, Sequencing of the sgRNA-targeted regions in the *SIRT6* gene, showing the coverage of the detected SIRT6 sequences in the wild-type (grey lines) and SIRT6^{−/−} (blue lines) monkeys. The red line highlights the SIRT6 sequence lost in SIRT6^{−/−} monkey no. 2, and the red triangles highlight the deletions (left, $\Delta 8$; right, $\Delta 27$) in SIRT6 in SIRT6^{−/−} monkey no. 1 and monkey no. 3. **d**, Western blots showing the absence of the SIRT6 protein and the alterations in H3K9ac and H3K56ac levels in the brains and muscles of SIRT6^{−/−} monkeys. For uncropped gels, see Supplementary Fig. 1.

¹National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. ²State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. ³University of Chinese Academy of Sciences, Beijing, China. ⁴Institute of Stem Cell and Regeneration, Chinese Academy of Sciences, Beijing, China. ⁵Advanced Innovation Center for Human Brain Protection, National Clinical Research Center for Geriatric Disorders, Cell Therapy Center, Xuanwu Hospital Capital Medical University, Beijing, China. ⁶Key Laboratory of Gene Engineering of the Ministry of Education, Department of Biochemistry, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. ⁷Department of Neurology, Peking University First Hospital, Beijing, China. ⁸These authors contributed equally: WeiQi Zhang, Haifeng Wan, Guihai Feng, Jing Qu. *e-mail: liwei@ioj.ac.cn; ghliu@ibp.ac.cn; byhu@ioj.ac.cn

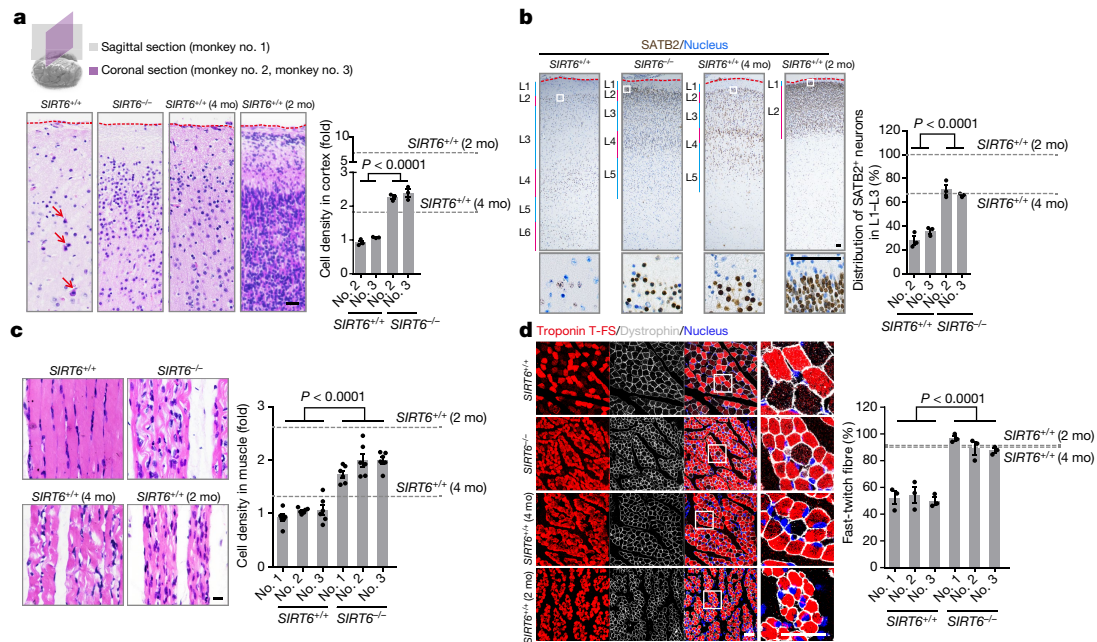


Fig. 2 | SIRT6 deficiency results in delays in the development of the brain and muscle. **a**, Schematic of the brain regions examined in our study (top). Haematoxylin and eosin staining, and calculation of the cell density, in cortical tissues from newborn wild-type and *SIRT6*^{-/-} monkeys, and from wild-type fetuses (bottom). The arrows identify pyramidal neurons. Scale bar, 100 μ m. **b**, Left, images of SATB2 immunostaining in the cortices of *SIRT6*^{-/-} and wild-type newborn, and wild-type fetal, monkeys. Right, the proportion of SATB2⁺ cells in L1–L3 cortical layers was quantified in brain sections. The blue and red lines represent different cortical layers. L, layer. Scale bars, 200 μ m. **c**, Haematoxylin and eosin staining (left) and calculated cell densities (right) in muscle tissues from *SIRT6*^{-/-} monkey, and wild-type newborn and fetal gastrocnemius and soleus muscles.

Scale bar, 50 μ m. **d**, Left, images of immunofluorescence staining for the fast-twitch marker troponin T-FS and sarcolemma-bound dystrophin in *SIRT6*^{-/-} muscles, and wild-type newborn and fetal gastrocnemius and soleus muscles. Right, the distribution of the different muscle fibre types is shown as a percentage of troponin T-FS⁺ fibres. Scale bar, 25 μ m. In **a**, **b**, the red dashed lines show the upper boundaries of the cortices. Grey dashed lines represent the average values for the 2-month-old and 4-month-old fetuses. White squares in **b**, **d** correspond to the enlarged area shown below (**b**) or to the right (**d**). $n = 3$ slices (**a**, **b**, **d**) or $n = 6$ slices (**c**), 3 images per slice, >200 cells per image. Data are mean \pm s.e.m., P values were determined by one-way ANOVA followed by Holm–Sidak’s multiple comparisons test.

Fig. 2a, b). Chromosomal variation analyses did not reveal any genomic or epigenomic instabilities in the tissues or mesenchymal stem cells of *SIRT6*-deficient monkeys (Extended Data Fig. 3a–k). Based on these results, we concluded that the *SIRT6* gene was successfully knocked out in the monkeys using the CRISPR–Cas9-based approach.

We noticed that *SIRT6*-null monkeys were substantially smaller at birth, weighing about 3.5 times less than wild-type newborn monkeys, but were similar in size to wild-type fetal monkeys with a gestational age of 2–4 months (Fig. 1a, b and Supplementary Table 1). Thus, the absence of *SIRT6* in the monkeys led to a whole-body developmental delay in utero and resulted in the birth of ‘premature’ offspring that failed to survive postnatally.

To determine whether the *SIRT6*-null monkeys also demonstrated developmental retardation at the tissue level, we first characterized the morphologies of the bone, fat, intestine epithelium, kidney, liver and lung. Compared to those of newborn wild-type monkeys, the morphological features of the *SIRT6*-null monkeys phenocopied those of wild-type fetuses—that is, lower bone density, the absence of subcutaneous fat, the appearance of immature intestine epithelium and higher cell density in various tissues (Extended Data Fig. 4a–f).

Brain and skeletal muscle are important organs that relate to energy and metabolic regulation. Compared with those of newborn wild-type monkeys, the brains of *SIRT6*^{-/-} monkeys were smaller and displayed thinner cortical layers in the cerebral and cerebellar hemispheres, resembling their wild-type fetal counterparts (Extended Data Figs. 5a–f, 6a–h). The brains of newborn *SIRT6*^{-/-} monkeys were also very similar to those of wild-type fetuses in terms of packing density, neuronal composition and layering of the cerebral cortex (Fig. 2a, b and Extended Data Figs. 5c, d, 6c, d), which may be related to the presence of immature HOPX⁺ outer-radial glia cells and a greater number of PAX6⁺ neural progenitor cells (NPCs) in the brains of

SIRT6^{-/-} monkeys (Extended Data Figs. 5e, 6e, f). Based on these data, it appears that *SIRT6* may function as a mediator of neural progenitor differentiation during monkey brain development, and that the absence of *SIRT6* delays neuronal maturation.

The muscle fibres in the *SIRT6*^{-/-} monkeys and wild-type fetuses displayed a similar density and size, and both differed from those of the muscles of newborn wild-type monkeys (Fig. 2c and Extended Data Fig. 7a). Wild-type muscle contained slow- and fast-twitch fibres, each of which exhibited distinct contractile and metabolic characteristics⁸ (Fig. 2d and Extended Data Fig. 7b–d), whereas muscles from the *SIRT6*^{-/-} monkeys mostly contained fast-twitch fibres that expressed a ubiquitous ATPase isoform and exhibited a high proportion of immature mitochondria, similar to the condition in fetal myoblasts⁸ (Fig. 2d and Extended Data Fig. 7b–d). These observations further support the hypothesis that an absence of *SIRT6* results in the developmental delay of many different tissues in monkeys.

To assess whether *SIRT6*^{-/-} monkeys sustain embryonic features at the transcriptome level, we performed RNA sequencing and compared the transcriptional profiles of the *SIRT6*^{-/-} monkeys with those of wild-type monkey fetuses. According to the principal component analysis, RNA profiling in the brains and muscles of the *SIRT6*-null monkeys closely correlated with RNA profiling in wild-type fetal brains and muscles; by contrast, brains and muscles of *SIRT6*-null monkeys exhibited a weaker correlation with the transcriptomic pattern in the corresponding tissues from wild-type newborns (Extended Data Fig. 8a). Consistent with these findings, a closer correlation was also observed between the global DNA methylation signatures of the brains of *SIRT6*^{-/-} newborns and of wild-type fetal brains (Extended Data Fig. 8b). Thus, the molecular developmental stage of the *SIRT6*^{-/-} monkeys mimics that of wild-type fetuses at 2–4 months of gestational age.

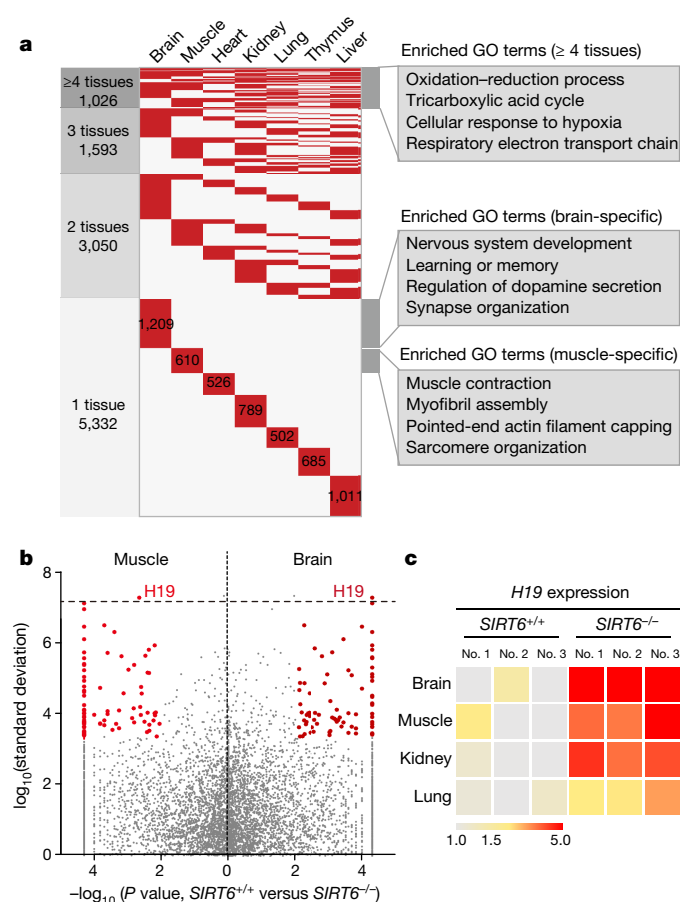


Fig. 3 | SIRT6 deficiency resulted in a developmental retardation at the transcriptome level. a, Heat map showing the differentially expressed genes in each organ from the *SIRT6*^{-/-} and *SIRT6*^{+/+} monkeys. The differentially expressed genes are coloured red, and genes that were not differentially expressed are depicted in light grey. The number of organs or tissues in which a gene was differentially expressed is shown in the left column. The enriched Gene Ontology terms for gene sets that were differentially expressed in at least four organs are shown; differentially expressed genes that were specifically expressed in the brain or muscle tissues are shown on the right. **b**, Scatter plots showing the top 200 variational developmental related genes used for the principal component analyses. Among these genes, differentially expressed genes (fold change (*SIRT6*^{-/-}/*SIRT6*^{+/+}) > 1.5, false-discovery-rate-adjusted *P* < 0.05) between wild-type and *SIRT6*-null brain and muscle are coloured red. The *P* value was calculated using Cuffdiff, through a one-sided *t*-test based on the Jensen–Shannon metric, under the null hypothesis of no change in relative abundances of genes. **c**, Heat maps showing normalized levels of *H19* expression in different samples. The average transcript levels in tissues from *SIRT6*^{+/+} monkey no. 1, monkey no. 2 and monkey no. 3 were normalized to 1, and the relative expression level of *H19* in each tissue was coloured as shown.

To characterize the molecular mechanisms that underlie the way in which *SIRT6* deficiency causes pan-tissue developmental retardation in monkeys, we compared genome-wide RNA expression profiles in seven organs from the *SIRT6*^{-/-} and wild-type newborn monkeys (Extended Data Fig. 8c–e and Supplementary Table 3). We identified 11,001 genes that were differentially expressed in at least one organ from the *SIRT6*^{-/-} and wild-type newborn monkeys. Among the differentially expressed genes, 1,026 were identified in at least 4 different organs. These genes are involved in aerobic respiration-related processes, such as the tricarboxylic acid cycle and the respiratory electron transport chain (Fig. 3a). These observations are consistent with previous reports that mice deficient in *SIRT6* exhibit alterations in glucose and energy metabolism, and with our observation that the

muscle cells of the *SIRT6*^{-/-} monkeys contain immature mitochondria^{1,6,9,10} (Extended Data Fig. 7d).

Among the differentially expressed genes, the long non-coding RNA *H19* was one of the most upregulated genes and was expressed in the brain at levels that were approximately 27.5-fold higher in *SIRT6*^{-/-} monkeys than in wild-type monkeys (Fig. 3b, c, Extended Data Fig. 8f and Supplementary Table 3). Consistent with the phenotypes, we also observed a higher level of the *H19* transcript in wild-type fetal monkeys than in their postnatal counterparts (Extended Data Fig. 8f). *H19* is a maternally expressed imprinted gene that regulates fetal development and has an integral role in the development of many tissues, including the brain^{11–13}. Importantly, excessive expression of *H19* in humans is linked to Silver–Russell dwarfism, a development retardation disorder characterized by intrauterine growth restriction^{12,13}. All these observations suggest that aberrant upregulation of *H19* may contribute to the developmental defects observed in *SIRT6*^{-/-} monkeys.

To further probe the roles of *SIRT6* in the brain and its conservation between monkey and human, we generated *SIRT6*-null human embryonic stem cells by a TALEN-mediated gene-editing procedure and then differentiated them into NPCs^{14,15} (Extended Data Fig. 9a). Both types of NPCs expressed the typical neural progenitor-specific markers PAX6, SOX2 and nestin, with no discernible differences in cell-cycle kinetics (Extended Data Fig. 9b–e). However, *SIRT6*^{-/-} NPCs displayed delayed neuronal differentiation (Fig. 4a). After two weeks of neuronal induction, PAX6, SOX2 and nestin were still expressed at higher levels in *SIRT6*^{-/-} cells than in wild-type cells (Fig. 4b). Furthermore, *H19* transcripts were more abundant in the human *SIRT6*-null neural cells than in wild-type cells (Fig. 4b), consistent with observations from the *SIRT6*-deficient monkey brain. Thus, we recapitulated *SIRT6*-dependent primate neuronal development using the in vitro human neuronal differentiation system.

We therefore investigated whether *H19* expression is directly controlled by *SIRT6* using chromatin immunoprecipitation (ChIP) followed by quantitative real-time PCR (qPCR). We observed an enrichment of *SIRT6* at the imprinting control region—located upstream of *H19* in the wild-type neuronal derivatives—but less enrichment was observed in wild-type NPCs (Fig. 4c). *SIRT6* specifically bound to this region, as *SIRT6* did not bind to adjacent loci (Fig. 4c and Supplementary Table 4).

Consistent with the increase in H3K56ac levels in the brains of the *SIRT6*-null monkeys and the wild-type fetuses, higher levels of H3K56ac were detected in the human *SIRT6*^{-/-} NPCs and neuronal derivatives (Extended Data Fig. 9e–g). Based on these results, we then investigated how the *SIRT6*–H3K56ac–*H19* axis regulates neuronal differentiation. In the absence of *SIRT6*, the imprinting control region of *H19* is occupied by more H3K56ac (Fig. 4d) and is accompanied by the recruitment of more CTCF—a trans-activator of *H19* transcription—to that region^{11,16–18} (Extended Data Fig. 9h). Ectopic overexpression of H3(K56Q), an H3K56 acetylation-mimic mutant¹⁴, in wild-type neural cells increased *H19* transcription and repressed neuronal differentiation (Extended Data Fig. 9i, j). Wild-type *SIRT6*, but not its catalytically inactive mutant (*SIRT6*(H133Y)), effectively rescued the changes in the levels of H3K56ac and *H19* in *SIRT6*-deficient cells (Extended Data Fig. 9i, k). Consistent with these findings, wild-type NPCs that overexpressed *H19* also displayed impaired neuronal differentiation (Fig. 4e and Extended Data Fig. 9l). By contrast, knockdown of *H19* in *SIRT6*^{-/-} NPCs improved the neuronal differentiation efficiency (Extended Data Fig. 9m). These findings supported the hypothesis that active *H19* is critical for maintaining the immature status of the *SIRT6*-null monkey brain.

In summary, our results not only highlight the possibility of generating genetically engineered monkeys that lack the longevity protein *SIRT6* across tissues, but also point to the fact that *SIRT6* deficiency results in global developmental delay in utero (Extended Data Fig. 10). Our study indicates that *SIRT6* functions as a mediator of primate brain development by repressing the expression of *H19* long non-coding RNA in a *trans* manner, a finding that adds to the complexity of *SIRT6*

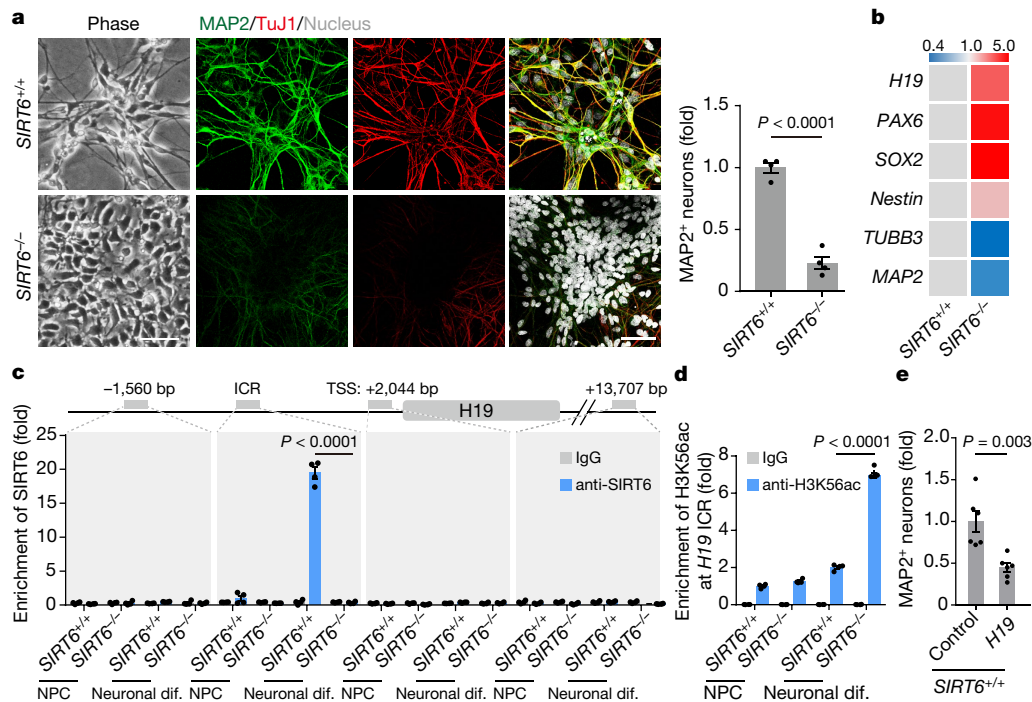


Fig. 4 | NPCs that lack SIRT6 exhibit arrested neuronal differentiation. **a**, Neuronal differentiation in wild-type and *SIRT6*^{-/-} NPCs. Left, images of immunostaining for the neuronal markers MAP2 and TuJ1. Right, quantification of MAP2⁺ neurons. *n* = 4 independent culture wells per condition. Scale bar, 50 μ m. **b**, Levels of neural genes in the *SIRT6*^{+/+} and *SIRT6*^{-/-} neuronal differentiation derivatives (neuronal dif.) were quantified using qPCR. The data for the *SIRT6*^{+/+} cells were used to normalize the corresponding data obtained from the *SIRT6*^{-/-} cells, and the relative expression levels of the genes in the *SIRT6*^{-/-} cells are coloured

from blue to red. *n* = 4 wells per condition. **c**, ChIP-qPCR assessment of the enrichment of SIRT6 at the imprinted control and adjacent control regions of *H19*. *n* = 4 wells per condition. **d**, ChIP-qPCR assessment of the enrichment of H3K56ac at the imprinted control region of *H19*. *n* = 4 wells per condition. **e**, Quantification of MAP2⁺ neurons during neuronal differentiation of wild-type NPCs that overexpress *H19* (see Extended Data Fig. 9). *n* = 6 wells per condition. Data are mean \pm s.e.m., *P* values were determined by two-sided Student's *t*-test (**a**, **b**, **e**) or two-way ANOVA (**c**, **d**). ICR, imprinted control region; TSS, transcription start site.

biology^{19,20}. Supporting our discovery, a loss-of-function mutation (D63H) of SIRT6 in humans was recently reported to cause late fetal loss with intrauterine growth restriction²⁰. Notably, human induced pluripotent stem cells derived from *SIRT6*(D63H) homozygous fetuses fail to differentiate into NPCs²⁰. The more severe phenotypes of *SIRT6*(D63H) fetuses and induced pluripotent stem cells may be attributed to the dominant negative effect of the *SIRT6*(D63H) mutant. The genetic monkey model presented here may open a new avenue through which to model and study the pathogenesis of human perinatal lethality syndrome.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0437-z>.

Received: 17 September 2017; Accepted: 16 July 2018;

Published online 22 August 2018.

- Houtkooper, R. H., Pirinen, E. & Auwerx, J. Sirtuins as regulators of metabolism and healthspan. *Nat. Rev. Mol. Cell Biol.* **13**, 225–238 (2012).
- Kanfi, Y. et al. The sirtuin SIRT6 regulates lifespan in male mice. *Nature* **483**, 218–221 (2012).
- Niu, Y. et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell* **156**, 836–843 (2014).
- Wan, H. et al. One-step generation of p53 gene biallelic mutant cynomolgus monkey via the CRISPR/Cas system. *Cell Res.* **25**, 258–261 (2015).
- Zuo, E. et al. One-step generation of complete gene knockout mice and monkeys by CRISPR/Cas9-mediated gene editing with multiple sgRNAs. *Cell Res.* **27**, 933–945 (2017).
- Chalkiadaki, A. & Guarente, L. Sirtuins mediate mammalian metabolic responses to nutrient availability. *Nat. Rev. Endocrinol.* **8**, 287–296 (2012).
- Kugel, S. et al. SIRT6 suppresses pancreatic cancer through control of Lin28b. *Cell* **165**, 1401–1415 (2016).
- Schiaffino, S. & Reggiani, C. Fiber types in mammalian skeletal muscles. *Physiol. Rev.* **91**, 1447–1531 (2011).

- Kugel, S. & Mostoslavsky, R. Chromatin and beyond: the multitasking roles for SIRT6. *Trends Biochem. Sci.* **39**, 72–81 (2014).
- Giblin, W., Skinner, M. E. & Lombard, D. B. Sirtuins: guardians of mammalian healthspan. *Trends Genet.* **30**, 271–286 (2014).
- Kernohan, K. D. et al. ATRX partners with cohesin and MeCP3 and contributes to developmental silencing of imprinted genes in the brain. *Dev. Brain* **18**, 191–202 (2010).
- Gabory, A., Jammes, H. & Dandolo, L. The *H19* locus: role of an imprinted non-coding RNA in growth and development. *BioEssays* **32**, 473–480 (2010).
- Wakeling, E. L. et al. Diagnosis and management of Silver-Russell syndrome: first international consensus statement. *Nat. Rev. Endocrinol.* **13**, 105–124 (2016).
- Pan, H. et al. SIRT6 safeguards human mesenchymal stem cells from oxidative stress by coactivating NRF2. *Cell Res.* **26**, 190–205 (2016).
- Liu, G. H. et al. Progressive degeneration of human neural stem cells caused by pathogenic LRRK2. *Nature* **491**, 603–607 (2012).
- Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).
- Kurukuti, S. et al. CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*. *Proc. Natl Acad. Sci. USA* **103**, 10684–10689 (2006).
- Fedorov, A. M., Stein, P., Svoboda, P., Schultz, R. M. & Bartolomei, M. S. Transgenic RNAi reveals essential function for CTCF in *H19* gene imprinting. *Science* **303**, 238–240 (2004).
- Schwer, B. et al. Neural sirtuin 6 (Sirt6) ablation attenuates somatic growth and causes obesity. *Proc. Natl Acad. Sci. USA* **107**, 21790–21794 (2010).
- Ferrer, C. M. et al. An inactivating mutation in the histone deacetylase SIRT6 causes human perinatal lethality. *Genes Dev.* **32**, 373–388 (2018).

Acknowledgements The authors acknowledge X. Wang, Y. Fu, L. Zhao, J. An and L. Wei for brain tissue analysis, S. Duan for western blotting, L. Sun from the Center for Biological Imaging for electron microscopy, R. Bai and P. Wang for technical assistance, L. Bai for administrative assistance, X. Liu for image processing using Vectra Automated Quantitative Pathology Imaging System, and P. Zhang for fluorescence in situ hybridization analysis. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA16010100), the National Key Research and Development Program of China (2015CB964800, 2017YFA0103304, 2017YFA0102802, 2014CB910503, 2014CB964600, 2018YFA0107203, 2016YFA0101403), the National Natural Science Foundation of China (91749202, 31471394, 31671429, 91749123, 81625009, 81330008,

81371342, 81471414, 81422017, 81601233, 81671377, 31601109, 31601158, 81771515, 81701388, 31571533, 31621004, 81422014), Program of Beijing Municipal Science and Technology Commission (Z181100001818002, Z151100003915072), Key Research Program of the Chinese Academy of Sciences (ZDRW-ZS-2017-5, KJZDEW-TZ-L05), Beijing Municipal Commission of Health and Family Planning (PXM2018_026283_000002) and Advanced Innovation Center for Human Brain Protection (117212).

Reviewer information *Nature* thanks H. Cohen, H. Yang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.-H.L., W.L., B.H. and Q.Z. conceptualized the work and supervised overall experiments; W.Z. performed the phenotypic and mechanistic analyses; H.W. generated gene-edited monkeys; G.F. performed bioinformatics analyses; J.Q. guided cell culture, differentiation and data analysis; Y.J. performed molecular experiments; R.R. performed brain biopsy; Z.C. and S.W. performed brain immunochemistry; Y.Y. and Z.W. performed

muscle analysis; Y.Z. performed fluorescence in situ hybridization; L.Z. designed sgRNAs; Z.L. and L.Z. performed genotyping; J.W. performed plasmid construction; G.-H.L., W.L., B.H., Q.Z., W.Z., H.W., G.F. and J.Q. performed data analysis and wrote the manuscript. All authors reviewed the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0437-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0437-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to W.L. or G.-H.L. or B.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Animals. All of the animal experiments were performed under the ethical guidelines of the Institute of Zoology, Chinese Academy of Sciences. Procedures using cynomolgus monkeys were approved by the Institutional Animal Care and Use Committee of Institute of Zoology, Chinese Academy of Sciences. All animals were housed at Beijing Institute of Xieixin Biology Resource with accreditation of Laboratory Animal Care accredited facility. Detailed information regarding the animals used in this study is included in Supplementary Table 1.

Oocyte recovery and embryo culture. The seven oocyte donors each received an intramuscular injection of recombinant human follitropin- α (GONAL-F, Merck Serono) twice daily for 8 days beginning on days 1–3 of their menstrual cycle. On day 9, the animals received an injection of recombinant human chorionic gonadotropin- α (OVIDREL, Merck Serono), and oocytes were aspirated laparoscopically 32 to 35 h later^{21,22}. The recovered MII-stage oocytes containing the first polar body were placed in hamster embryo culture medium 10 (HECM-10) before intracytoplasmic sperm injection, after which the fertilized embryos were each cultured in 50 μ l of HECM-10 containing 10% (v/v) fetal bovine serum. **sgRNA and DNA plasmid construction.** Six sgRNAs with different sequences that target conserved domains in *SIRT6* exon 5 (2 sgRNAs), exon 7 (1 sgRNA) or exon 8 (3 sgRNAs) were designed using established protocols^{23,24}. Each synthetic 24-bp oligonucleotide containing a 20-bp spacer sequence and a 4-bp overhang that had been annealed to form a duplex was cloned into the pUC19-U6-sgRNA scaffold at the *BbsI* site.

The full-length *H19* long non-coding RNA was synthesized and cloned into a pHIV lentiviral vector (BGI). The lentiviral expression vectors were generated by cloning the respective cDNA into the PLE4 vector (a gift from T. Hishida). For the construction of the lentivirus vectors expressing the short hairpin RNA targeting *H19* (Supplementary Table 4), the corresponding short hairpin RNA oligonucleotides were inserted into the multiple cloning site of pLVTHM (Addgene, 12247). Lentivirus particles were produced in transfected HEK293T cells and used to transduce NPCs in the presence of 10 μ M ROCK inhibitor and 4 μ g/ml polybrene.

Targeting efficiencies of the sgRNAs. For each sgRNA, a Cas9 plasmid (3 μ g) and one sgRNA plasmid (2 μ g) were transfected into monkey Vero cells. Cells that were positive for green fluorescent protein were sorted by FACS 48 to 72 h later (Beckman Coulter, MoFlo XDP). The transfection efficiency of each sgRNA was determined by PCR and a T7 endonuclease I assay, followed by Sanger sequencing.

Generation of *SIRT6*^{-/-} monkeys. The RNAs used for microinjection were transcribed in vitro. The Cas9 mRNA and sgRNA were transcribed in vitro using reagents from the HiScribe T7 High Yield RNA Synthesis kit (NEB) according to the manufacturer's instructions. The m7G(5')ppp(5')G RNA Cap Structure Analogue reagent was used to cap the in vitro transcribed Cas9 mRNA. After the capping reaction, *Escherichia coli* poly(A) polymerase (NEB) was used to add a 3'-poly(A) tail, which stabilized the Cas9 mRNA. The transcribed Cas9 mRNA and the sgRNAs were purified using MicroElute RNA kit reagents (Omega).

Ninety-eight zygotes, each with a clear pronucleus, were selected for intracytoplasmic injection. Under a Leica standard microinjection system, a programmable microinjector (Femto Jet, Eppendorf) injected 4 pl of a solution containing the transcribed RNAs into each zygote, which were then cultured in HECM-10 supplemented with 10% (v/v) fetal calf serum until their transfer into the surrogate mothers. Forty-eight high-quality embryos at the 2- to 16-cell stages were transferred into the oviducts of the 12 matched recipient monkeys, with 3–5 embryos transferred per surrogate. The earliest stage of pregnancy, as defined by fetal cardiac activity and the presence of a yolk sac, was detected by ultrasonography between day 20 and 30 after embryo transfer.

Genotyping and sequencing. Genomic DNA was extracted from organs, including the brain, muscle, heart, liver and lung, from wild-type and *SIRT6*-null monkeys. Then, the targeted fragments of the *SIRT6* gene were amplified using PrimerSTAR polymerase (TaKaRa) with the primers shown in Supplementary Table 4. The amplified fragments were subcloned into pEASY-Blunt (TransGen) for sequencing.

Western blotting. Monkey tissues were homogenized in liquid nitrogen, and proteins extracted with TRIzol reagent (Invitrogen). Western blotting was performed using previously described methods¹⁵. Protein concentrations were quantified using a BCA kit. Proteins (20 μ g per lane) were electrophoresed through an SDS-PAGE gel and subsequently electro-transferred to a polyvinylidene fluoride membrane (Millipore). Antibodies used in the present study were anti-*SIRT6* (Cell Signaling Technology, 12486, 1:1,000 dilution), anti-H3K56ac (Abcam, ab76307, 1:5,000 dilution), anti-H3K9ac (Millipore, 06-942, 1:1,000 dilution), anti-H3K9me3 (Abcam, ab8896, 1:2,000 dilution), anti-HP1 α (Cell Signaling Technology, 2616, 1:2,000 dilution), anti-H3 (Santa Cruz Biotechnology, sc-10809, 1:1,000 dilution), anti-Flag (Sigma, F2555, 1:5,000 dilution) and anti-GAPDH (Abcam, ab9485, 1:5,000 dilution). Secondary antibodies from CST were used to

visualize the proteins. All the raw data of the blots are provided in Supplementary Fig. 1.

qPCR. Total RNA was extracted with TRIzol reagent, according to the manufacturer's instructions. For cDNA synthesis, 2 μ g of total RNA was reverse transcribed with Reverse Transcription Master Mix (Promega). qPCR was performed using iTaq Universal SYBR Green Supermix (Bio-Rad), according to the manufacturer's instructions. The expression of cDNAs in each sample was normalized to GAPDH. Telomere lengths were determined by qPCR²⁵. Primer sequences for the experiment are listed in Supplementary Table 4.

DNA extraction, library construction and sequencing. Genomic DNA was extracted from the cortices of the *SIRT6*^{-/-} and wild-type monkeys using the DNeasy Blood & Tissue kit (Qiagen), according to the manufacturer's instructions. DNA was randomly fragmented into ~500 bp lengths using a Covaris ultrasonic processor. Fragmented DNA was ligated with DNA adaptors (Illumina) and amplified using Illumina paired-end PCR primers. DNA libraries were quantified using a Qubit 2.0 Fluorometer (Life Technologies) and analysed using an Agilent Bioanalyzer 2100 to determine the insert sizes of the fragments in the libraries. The library of fragments was sequenced on an Illumina HiSeq X Ten platform.

Bioinformatics analyses of single nucleotide variants, copy-number variations and repeated sequences. The resequenced reads of the *SIRT6*^{-/-} and wild-type DNA samples were mapped to the *M. fascicularis* reference genome (macFas5) using a Burrows–Wheeler Aligner (BWA, version 0.7.15). Duplicate reads were removed and the uniquely mapped reads were retained for the copy-number variation (CNV) analysis, in which chromosomal sequences were placed into bins of 500 kb in length. The normalized coverage depth for each bin was calculated by dividing the raw coverage depth by the average sequencing depth. The repeat regions annotated for *M. fascicularis* by RepeatMasker (db20140131) (<http://www.repeatmasker.org>) were removed from the genomic sequence before calculating the coverage. The CNV scatterplot was drawn using ggplot2. The read coverage of *SIRT6* genome region is shown in Fig. 1c and produced by IGV (version 2.3.88)²⁶.

For the single nucleotide variants (SNV) analysis, the read base sites with an incorrect base probability >0.001 were replaced with an N. The base distribution for each chromosomal location was calculated using the pysamstats (version 1.0.0) (<https://github.com/alimanfoo/pysamstats>). Owing to the poor quality of the monkey chromosome assembly, only the masked chromosome segments conserved between the human (hg19) and the monkey (macFas5) were used for the SNV analysis. The predominantly proportional base in each site was used as the reference base to avoid including unannotated single nucleotide polymorphisms (SNPs). The heterozygosity of each site was calculated as the depth of the enriched second base divided by the reference base depth. The base heterozygosity defined a SNP site. A site with a heterozygosity percentage >0% and <30% was defined as an SNV site induced by random mutation.

For the repeat-sequence analysis, Repbase (v.21.11) annotated repeat sequences were used to construct the reference sequence index. Reads were mapped to the indexed repeat sequences, and the mapped reads were grouped into long or short interspersed elements (LINEs or SINEs, respectively), long terminal repeats (LTRs), ribosomal RNAs (rRNAs) and other types of repeats. The number of reads mapped to each type of repeat was normalized to the total sequencing depth.

Off-target site assay. The whole-genome sequencing data were used to analyse off-target CRISPR–Cas9 editing. As previously described²⁷, the whole-genome sequencing data were mapped to macFas5 genome by BWA (version 0.7.15) using 'mem' mode with default parameters. The mapped reads with unique genome location were used for the indel identification. The reads in mapped .sam files with CIGAR value 'D', 'I' or 'N' were considered as potential indel-containing reads. Pysamstats (version 1.0.0) under default settings was used to extract these potential indel sites with two reads supporting and no less than 10 \times coverage for the next analysis. First, the sites existing in one of the wild-type genomic sequencing datasets were removed. Next, candidate indels located in repeats and low-complexity regions annotated by RepeatMasker (db20140131) were filtered. The homologous sites annotated as indel-type SNPs in humans and mice were also discarded. Then, if one homopolymer larger than 5 bp or two 4-bp homopolymers as well as three 3-bp homopolymers were present in regions located near the indel site (30 bp each, up and downstream of the indel site), the indel was also discarded. In the same region, the unannotated low-complexity regions with a single same base content >40% were also removed. Finally, regardless of the similarity to sgRNA sequences, the indel sites located within the protein-coding regions of genes annotated in the Ensembl genome database were considered as possible functional sites. Thus, 14, 16 and 6 indel sites were identified in the three *SIRT6*^{-/-} samples, respectively. No overlapping sites were identified in all three samples, as shown in Extended Data Fig. 1h.

Simultaneously, all potential off-target sites with homology to the 20-bp sequences (sgRNA plus protospacer adjacent motif (PAM) sequences) were retrieved from the *M. fascicularis* genome (macFas5). Possible off-target sites

with no more than five mismatched sites were identified using Cas-OFFinder²⁸. In addition, 3,891 possible off-target regions were identified, including 983 sites without NAG or NGG PAM sequences. None of these regions contained indel sites identified by genomic sequencing, even within the non-coding sequence (CDS) regions, such as introns, untranslated regions (UTRs) or intergenic regions. The sequencing read coverage for the possible off-target sites in the CDS region were also checked, as shown in Extended Data Fig. 1g. The genomic sequencing read coverage for the possible off-target regions (10-bp up- and downstream, as well as the 20-bp homology region) was calculated for all three *SIRT6*^{-/-} and wild-type samples. The score for each site of all regions was calculated as $-\log_2((\text{coverage in } SIRT6^{-/-} \text{ sample})/(\text{average coverage in three wild-type samples}))$. The highest site scores for the region were used as the region score, as shown in Extended Data Fig. 1g. The results did not reveal notable sequencing coverage changes in these regions, consistent with the indel results. For the two injected sgRNAs that target *SIRT6*, we predicted four and three possible off-target sites in CDS regions of all annotated genes. Then, we used the resequenced genomic data to identify single-base mutations, insertions or deletions in these sites.

Karyotyping analysis and fluorescence in situ hybridization. A standard G-banding analysis and fluorescence in situ hybridization were performed on spread chromosomes using previously published protocols²⁹.

RNA-seq library construction and data analysis. Total RNA was extracted using TRIzol reagent, as described above, and purified using an RNeasy Mini kit (Qiagen), according to the manufacturer's instructions. After quantification on a Fragment Analyzer (Advanced Analytical), RNA (1.5 µg) and a TruSeq RNA Sample Preparation kit (Illumina) were used to construct the sequencing libraries according to the manufacturer's standard protocol. Libraries were sequenced on an Illumina platform (HiSeq X Ten). Clean reads were mapped to the macFas5 genome using HISAT version 0.1.6-beta³⁰. Each read with a unique genomic location in the bam format was retained for a gene expression calculation using Cufflinks (version 2.0.2) with the option '-GTF³¹'. The gene annotation was downloaded from NCBI (GCF_000364345.1), and only exon annotations for mRNAs and other non-coding RNAs were used. Genes with at least 1 fragment per kilobase of exon per million fragments mapped (FPKM) in at least 1 sample were used to analyse the differentially expressed genes (fold change (*SIRT6*^{-/-}/*SIRT6*^{+/+}) > 1.5, false-discovery-rate-adjusted $P < 0.05$). Gene Ontology terms for the differentially expressed genes were assigned by DAVID (version 6.7), and biological processes were selected based on P values < 0.05. Figures were produced using ggplot2 (version 2.2.1)^{32,33}. The three criteria used to confirm our differentially expressed genes results were reliable. First, the P value was calculated by Cuffdiff on the basis of the square root of the Jensen-Shannon divergence to evaluate the difference compared to corresponding wild-type samples³⁴. Second, for P values less than 0.05, we also set a stricter FDR of less than 0.05 by performing multiple tests to reduce the chance of type I errors when testing the null hypothesis. Third, only genes with no less than 1 FPKM in at least 1 sample were used to analyse the differentially expressed genes and avoid false positive results produced by genes expressed at low levels.

The top 200 variational developmental (GO:0032502) related genes that were ranked by standard deviation among all brain and muscle samples were used for the principal component analysis. All expressed genes (no less than 1 FPKM in at least 1 sample) were used for t -distributed stochastic neighbour embedding (t -SNE) analysis. The principal component analysis and the t -SNE analyses were performed using prcomp (basic function in R version 3.2.5) and Rtsne (version 0.11) function in the R package, respectively. The heat maps and other related figures were produced using heatmap.2 (function in package gplots version 3.0.1) or ggplot2 (version 2.2.1), respectively.

Whole-genome bisulfite sequencing. Genomic DNA from tissues and cells was used to construct the library according to the Illumina protocol, with minor modifications. Unmethylated lambda-DNA (Promega) was spiked into the lysis buffer to track the efficiency of bisulfite conversion. Bisulfite conversion of the adaptor-ligated DNA fragments was performed using the EZ DNA Methylation-Gold Kit (Zymo Research) according to the manufacturer's protocol. Sequencing libraries were prepared using KAPA HiFi HotStart Uracil + ReadyMix (2×) and were sequenced on the HiSeq X Ten platform (sequenced by Novogene). Raw sequences obtained from whole-genome bisulfite sequencing were examined for quality and the bisulfite conversion rate. Subsequently, all monkey tissues were analysed using the same pipeline. In brief, the sequencing reads were trimmed by Trim Galore (version 0.4.2) (Babraham Bioinformatics) and mapped to the monkey (macFas5) using Bismark v.0.13.1 (Babraham Bioinformatics). The methylation levels of covered cytosine sites were calculated by dividing the number of reported C sites by the total number of reported C and T sites. CpG sites covered by more than ten reads in monkeys were used for the analysis. The cluster analysis of global DNA methylation in monkey brain samples was performed on the promoter region (2-kb upstream and 1-kb downstream of the TSS site) of 1,607 development-related genes, which predominantly contributed to the observed differences in expression

among *SIRT6*^{+/+} newborn monkeys and *SIRT6*^{-/-} monkeys in the RNA sequencing analysis. Only CpG sites displaying variations in the methylation level greater than 50 among the samples were used for the cluster analysis. The cluster analysis was performed using the cluster function in R.

Histology and immunofluorescence staining. For the histological analysis, tissues such as the brains, muscles and kidney were collected, fixed with paraformaldehyde and embedded in paraffin. Paraffin-embedded tissue sections were stained with haematoxylin and eosin or used for immunohistochemistry according to standard methods³⁵. Left hemispheres of each brain were stored in liquid nitrogen for further analysis, and right hemispheres were fixed and sectioned into frozen sagittal slices (*SIRT6*^{+/+} (no. 1) and *SIRT6*^{-/-} (no. 1)) or paraffin coronal slices (*SIRT6*^{+/+} (no. 2/no. 3) and *SIRT6*^{-/-} (no. 2/no. 3)), according to the expertise of the neuropathologists. For immunohistochemistry, paraffin-embedded tissue sections were subjected to a heat-mediated antigen retrieval procedure, and then endogenous peroxidases were blocked with the Vector Laboratories Bloxall dual-enzyme blocking solution. Next, tissue sections were incubated with a primary antibody overnight. Finally, the appropriate ImmPRESS Reagent (Vector Laboratories) was added to the sections, which were then incubated for 30 min. Antigen-positive cells were visualized using the ImmPACT DAB Substrate kit (Vector Laboratories). Primary antibodies included: anti-H3K56ac (Abcam, ab76307, 1:1,500 dilution), anti-SATB2 (Abcam, ab92446, 1:250 dilution), anti-NeuN (Millipore, MAB377, 1:200 dilution), anti-PAX6 (COVANCE, PRB-278P, 1:200 dilution), anti-PCP2 (Santa Cruz Biotechnology, sc-137064, 1:200 dilution), anti-BRN2 (Santa Cruz Biotechnology, sc-6029, 1:200 dilution), and anti-HOPX (Santa Cruz Biotechnology, sc-30216, 1:200 dilution). For each antibody, negative staining with IgG controls was performed to ensure specificity. Images were captured using a Vectra Automated Quantitative Pathology Imaging System (Perkin Elmer) or an Olympus DP72 microscope. For the quantification of neurons, brains were dissected and the slices with lateral geniculate nucleus—as defined by a certified pathologist—were stained with haematoxylin and eosin. HOPX⁺ and SATB2⁺ neurons were counted in the cortex, PAX6⁺ NPCs were counted in the hippocampus and PCP2⁺ neurons were counted in the cerebellum.

For ATPase staining, gastrocnemius and soleus muscles were snap-frozen in liquid nitrogen, cryosectioned and stained for ATPase activity using a standard procedure³⁶. Slow-twitch fibres were lightly stained at pH 10.6 and intensely stained at pH 4.4. Fast-twitch fibres were intensely stained at pH 10.6 and lightly stained at pH 4.4. The numbers of positively stained fibres were counted and reported as a percentage of the total number of fibres.

Immunofluorescence staining was performed as previously described³⁵. Muscles were frozen and sectioned into slices using a freezing microtome (Leica). Frozen sections of brain or cultured cells were fixed with 4% paraformaldehyde and were subjected to immunofluorescence staining using an established protocol³⁵. The primary antibodies were anti-troponin T-FS (Santa Cruz Biotechnology, sc-166663, 1:200 dilution), anti-dystrophin (Abcam, ab15277, 1:200 dilution), anti-myosin-slow (Sigma, M8421, 1:200 dilution), anti-H3K56ac (Abcam, ab76307, 1:1,000 dilution), anti-MAP2 (Sigma, M4403, 1:500 dilution), anti-TuJ1 (Sigma, T8578, 1:500 dilution), anti-Ki67 (Vector Laboratories, VP-RM04, 1:500 dilution), anti-SOX2 (Santa Cruz Biotechnology, sc-17320, 1:200 dilution), anti-nestin (Millipore, MAP5326, 1:500 dilution), and anti-PAX6 (COVANCE, PRB-278P, 1:200 dilution). Images were visualized and photographed using a Leica SP5 confocal microscope. Quantitative analyses were performed on more than 200 randomly selected cells from each sample using ImageJ software.

For the images of histological and immunofluorescence staining, brain sections from wild-type and *SIRT6*-null monkeys were subjected to serial slicing, and one slice every ~100 or ~50 slices was stained, and at least three different tissue layers were characterized and quantified in each group.

Bone mineral density measurements. Bone mineral density was measured using a Quantum FX microCT system (Perkin Elmer).

Cell culture. H9 human embryonic stem cells (WiCell Research) and their *SIRT6*-null derivatives were maintained on mitomycin-C-treated mouse embryonic fibroblasts in human embryonic stem cell medium or on Matrigel-coated plates (BD Biosciences) in mTeSR medium (STEMCELL Technology). The embryonic stem cell medium contained DMEM/F12 (Invitrogen) supplemented with 20% (v/v) Knockout Serum Replacement (Invitrogen), 1 × non-essential amino acids (Invitrogen), 1 × GlutaMAX (Invitrogen), 55 µM β-mercaptoethanol (Invitrogen) and 10 ng/ml bFGF (Joint Protein Central). HEK 293T cells were maintained in high-glucose DMEM (Invitrogen) supplemented with 10% (v/v) FBS. Mesenchymal stem cells were successfully isolated from the bone marrow of monkey no. 2 and no. 3, purified by FACS sorting (CD73 and CD105 double positive), and cultured in 90% α-MEM + 10% FBS (Gibco) supplemented with Glutamax (Gibco), 1% penicillin/streptomycin (Gibco) and 1 ng/ml FGF2 (Joint Protein Central). 293T was authenticated by ATCC. H9 human embryonic stem cells were authenticated by WiCell. The properties of all stem cells were

authenticated by karyotyping. All cell lines tested negative for mycoplasma contamination.

Generating NPCs from embryonic stem cells. NPC induction was performed as previously described¹⁵. In brief, embryonic stem cells were cultured on mouse embryonic fibroblasts, and when they reached 70–80% confluence, cells were incubated with neural differentiation medium-1 composed of 50% (v/v) advanced DMEM/F12 (Invitrogen), 50% (v/v) Neurobasal medium (Invitrogen), 1 × N2 supplement (Invitrogen), 1 × B27 supplement (Invitrogen), 1 × GlutaMAX, 10 ng/ml hLIF (Millipore), 2 μM dorsomorphin, 3 μM SB431542 (Tocris), 4 μM CHIR99021 (Stemgent) and 0.1 μM compound E (EMD Chemicals). After 2 days, the medium was replaced with fresh medium, and cells were cultured for an additional 5 days. Cultures were then dissociated into single cells with Accumax (Innovative Cell Technologies), transferred to Matrigel-coated plates, and maintained in Neural Progenitor Cell Maintenance Medium (50% (v/v) advanced DMEM/F12, 50% (v/v) Neurobasal medium, 1 × N2 supplement, 1 × B27 supplement, 2 mM GlutaMAX, 10 ng/ml hLIF, 2 μM SB431542 and 3 μM CHIR99021). **Neuronal differentiation.** NPCs (5,000 per well) were cultured on Matrigel-coated 24-well plates for 3 days and then neuronal differentiation was induced by incubating cells in DMEM/F12 medium containing 1 × N2 supplement, 1 × B27 supplement, 200 μM ascorbic acid (Sigma), 400 μM dbcAMP (Sigma), 10 ng/ml GDNF (Peprotech) and 10 ng/ml BDNF (Peprotech). Laminin (Sigma) was added to the cultures after 2 days to facilitate differentiation. Cells were maintained in the aforementioned medium for 14 days and then immunostained for the neuronal markers MAP2 and TuJ1.

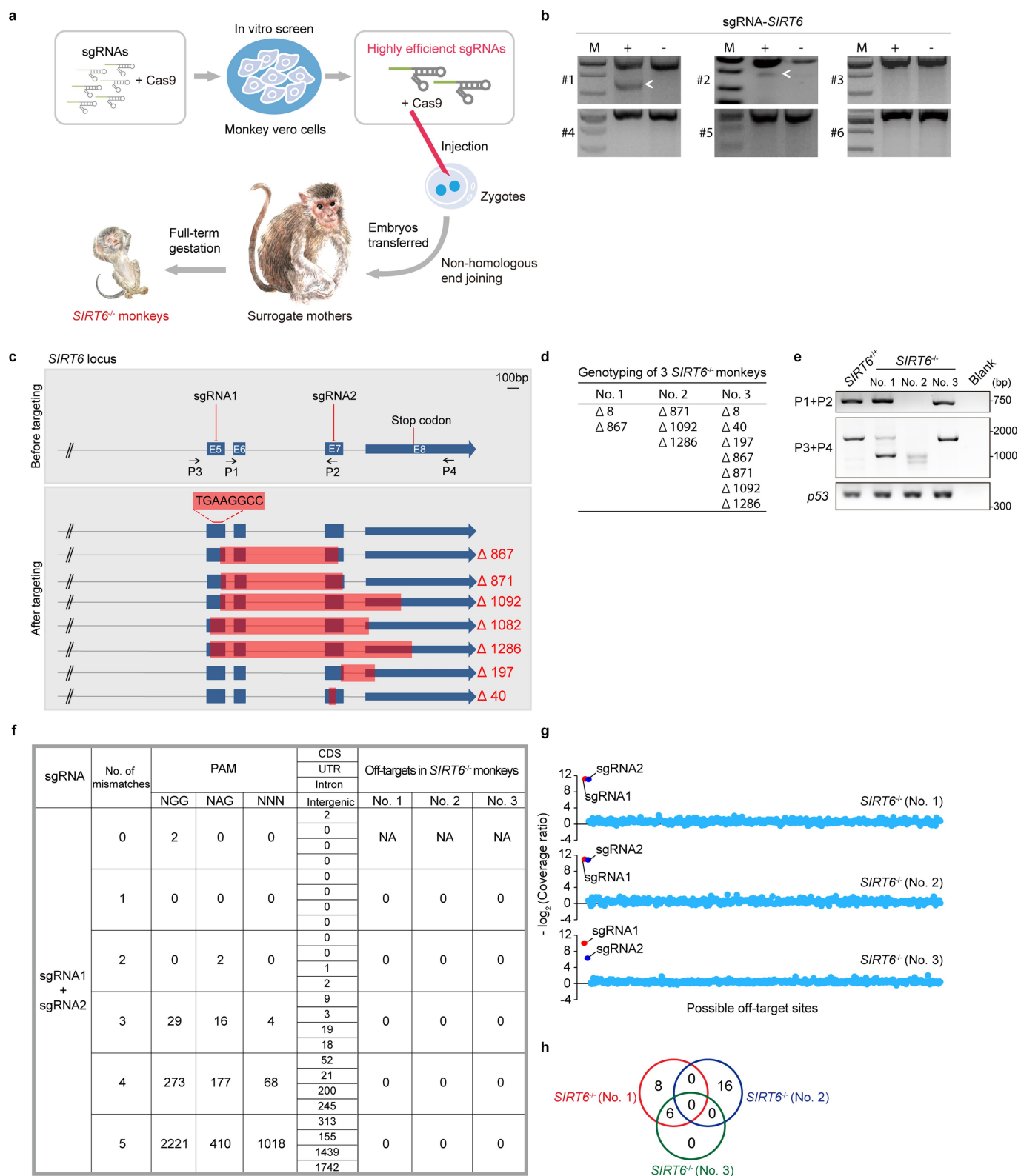
ChIP. ChIP was performed as previously described²⁵, with minor modifications. In brief, cells were cross-linked with 1% w/v formaldehyde (Sigma) for 10 min at room temperature, and then the crosslinking was quenched by incubating the cells with 125 mM glycine for 5 min at room temperature. Next, 2 × 10⁶ cells were lysed in buffer containing 50 mM Tris-HCl, 10 mM EDTA and 1% (w/v) SDS, pH 8.0, and the chromatin from the cells was sheared using a Covaris S2 instrument. Chromatin was then incubated overnight (12–16 h) with samples of protein A Dynabeads bound to 2.4 μg of anti-SIRT6 (CST, 12486), 1.2 μg of anti-H3K56ac (Abcam, ab76307) or 2.4 μg of anti-CTCF (Millipore, 07-729) antibodies. An equal amount of rabbit IgG (CST) was used as the negative control. Antibody-tagged chromatin was eluted with 20 mM Tris-HCl, 5 mM EDTA and 50 mM NaCl, pH 7.5, and then incubated with proteinase K on a thermomixer at 1,300 rpm for 2 h to obtain the DNA of interest. Next, the DNA was purified by phenol-chloroform-isoamyl alcohol extraction, precipitated with ethanol, and subjected to qPCR with the appropriate primers on the imprinting control region or adjacent loci (Supplementary Table 4). Primers for the imprinting control region of *H19* (1977613–1977821, NCBI Build 36) and adjacent control loci on chromosome 11 were designed according to previous studies^{37,38}.

Statistical analysis. Data were analysed using two-tailed Student's *t*-tests and are presented as means ± standard errors of the means. For multiple comparisons, the *P* value was calculated using GraphPad Prism software by two-way ANOVA or one-way ANOVA, followed by the recommended Holm–Sidak method. Cell density was determined by counting the nuclei per unit area, the cortical thickness and the diameters of HOPX⁺ and PCP2⁺ cells were measured using ImageJ software.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All relevant data are available from the authors and the Source Data represented graphically in the figures are available with the paper. The sequencing data have been deposited in the NCBI Gene Expression Omnibus (GEO) under the accession number GSE102830, NCBI Sequence Read Archive under accession number SRP149748 and Genome Sequence Archive of Beijing Institute of Genomics, Chinese Academy of Sciences (<http://gsa.big.ac.cn/>) under accession number PRJCA000909.

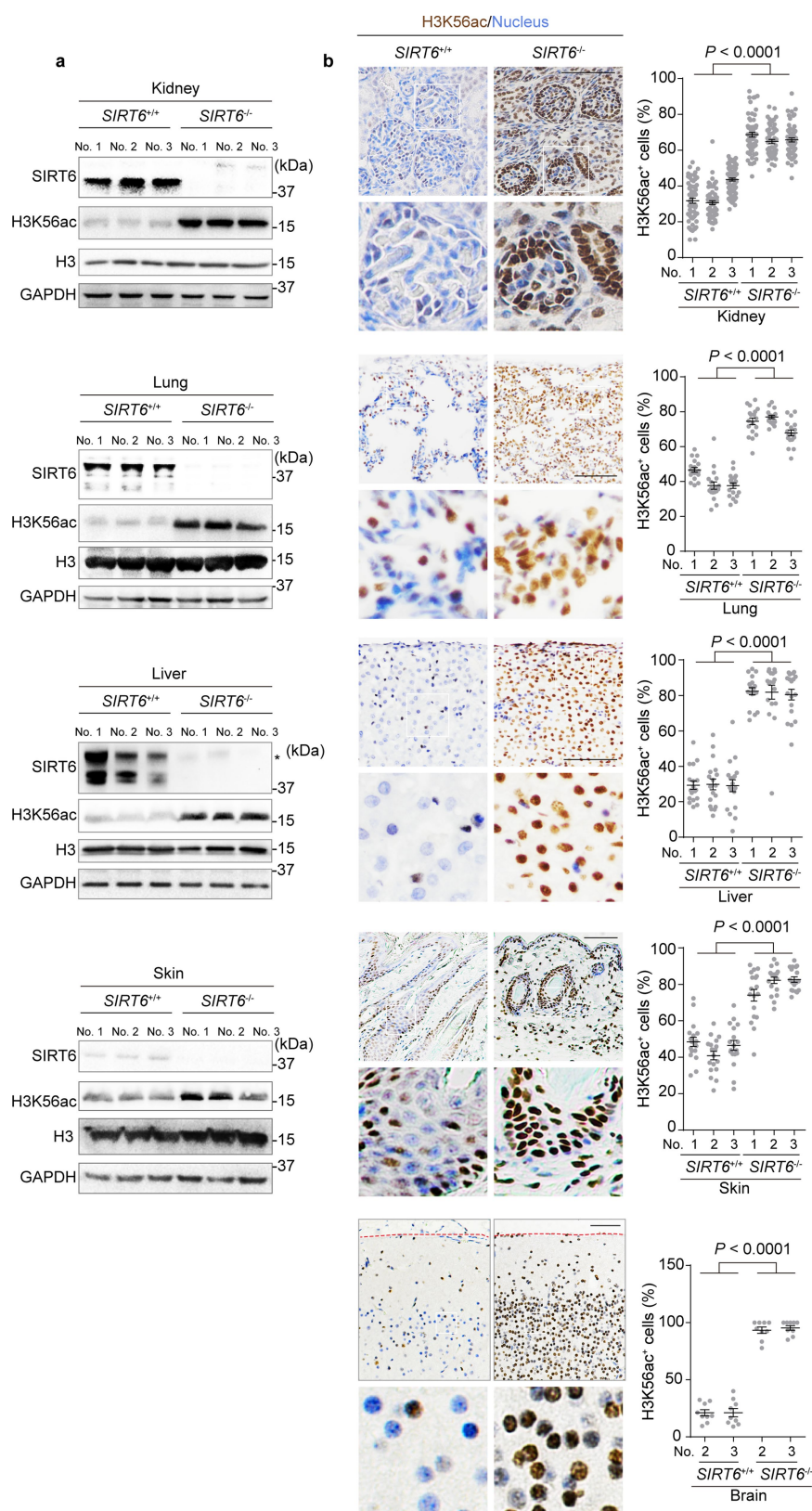
21. Yang, S. et al. Effects of rhFSH dose on ovarian follicular response, oocyte recovery and embryo development in rhesus monkeys. *Theriogenology* **67**, 1194–1201 (2007).
22. Niu, Y. et al. Transgenic rhesus monkeys produced by gene transfer into early-cleavage-stage embryos using a simian immunodeficiency virus-based vector. *Proc. Natl Acad. Sci. USA* **107**, 17663–17667 (2010).
23. Wang, H. et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
24. Pyzocha, N. K., Ran, F. A., Hsu, P. D. & Zhang, F. RNA-guided genome editing of mammalian cells. *Methods Mol. Biol.* **1114**, 269–277 (2014).
25. Zhang, W. et al. A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science* **348**, 1160–1163 (2015).
26. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
27. Veres, A. et al. Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. *Cell Stem Cell* **15**, 27–30 (2014).
28. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
29. Mao, P. et al. Homologous recombination-dependent repair of telomeric DSBs in proliferating human cells. *Nat. Commun.* **7**, 12154 (2016).
30. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
31. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
32. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
33. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
34. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
35. Ocampo, A. et al. In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell* **167**, 1719–1733.e12 (2016).
36. Chalkiadaki, A., Igarashi, M., Nasamu, A. S., Knezevic, J. & Guarente, L. Muscle-specific SIRT1 gain-of-function increases slow-twitch fibers and ameliorates pathophysiology in a mouse model of duchenne muscular dystrophy. *PLoS Genet.* **10**, e1004490 (2014).
37. Nativo, R. et al. Cohesin is required for higher-order chromatin conformation at the imprinted *IGF2-H19* locus. *PLoS Genet.* **5**, e1000739 (2009).
38. Kim, H. S. et al. The hSsu72 phosphatase is a cohesin-binding protein that regulates the resolution of sister chromatid arm cohesion. *EMBO J.* **29**, 3544–3557 (2010).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Generation and genotyping of *SIRT6*^{-/-} monkeys. **a**, Schematic showing the optimization of the CRISPR–Cas9-based targeting system for *SIRT6* in monkeys. **b**, T7 endonuclease I assay for determining the indel rates of the six sgRNAs that target the monkey *SIRT6* gene. The white arrows indicate the cleavage bands. M, marker. **c**, Schematic of the mosaic mutations in the *SIRT6* loci. The red blocks identify the nucleotide sequences deleted in the *SIRT6* gene. E, exon. **d**, Types of deletions identified by Sanger sequencing of the PCR-amplified *SIRT6* cDNA from organs of the three *SIRT6*^{-/-} monkeys. **e**, PCR of the *SIRT6* gene from the wild-type and *SIRT6*^{-/-} monkey genomes using primer pairs P1 and P2 or P3 and P4 (Supplementary Table 4). PCRs of the p53 locus were used as a control. **f**, Whole-genome sequencing data did not reveal mutations in the potential off-target sites predicted on the

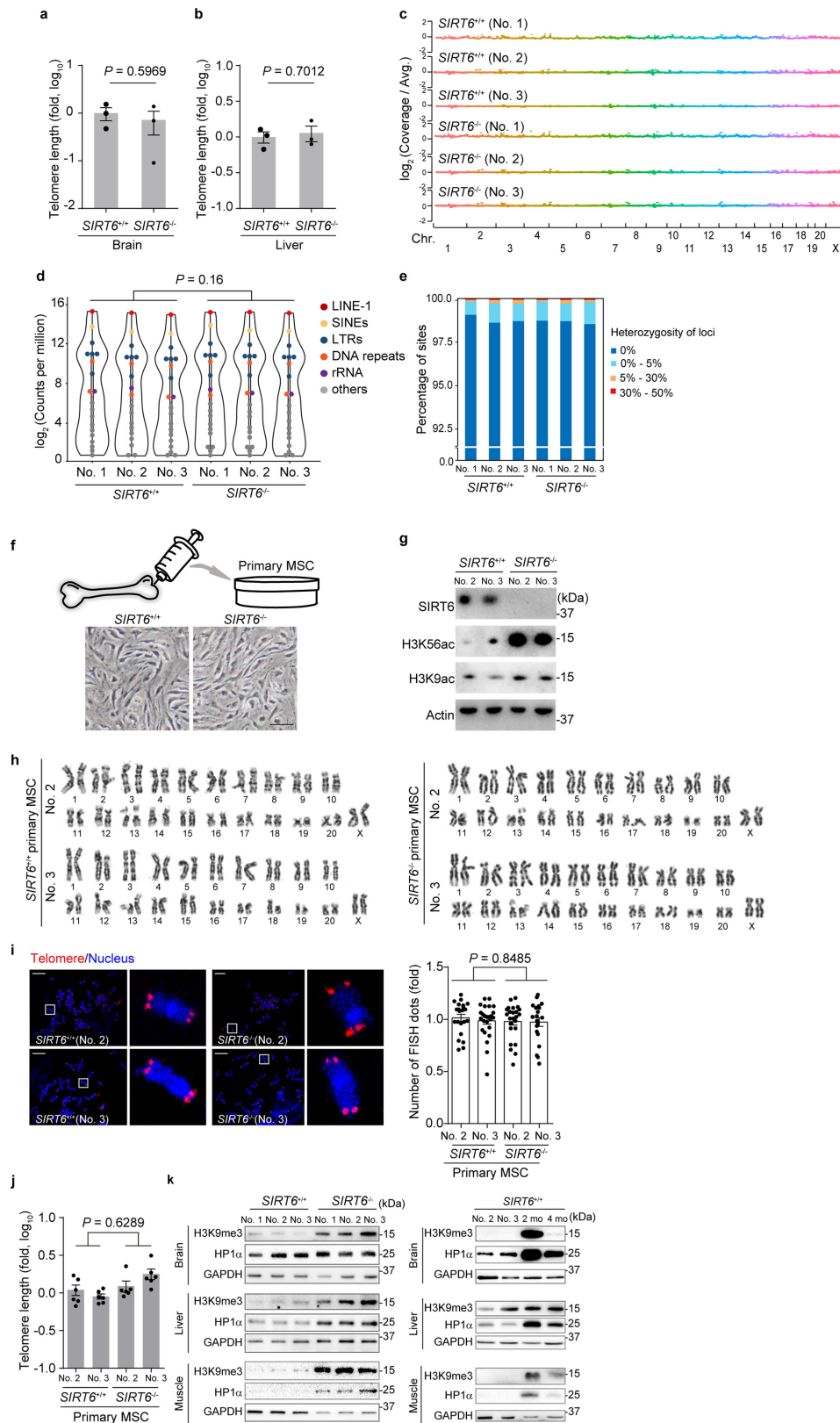
basis of sequences similar to the sgRNA sequence. The presumed PAM sequences and genomic locations of these sites are shown. **g**, Scatter plots showing the coverage scores for targeting sites within the *SIRT6* gene and predicted potential off-target sites for the two sgRNAs. Three hundred and seventy-four predicted potential off-target sites and two *SIRT6* sgRNA-target sequences located in the CDS regions are presented. The methods for calculating the coverage score are described in Methods. **h**, Venn diagram showing the lack of consistency of mutations in CDS regions detected by whole-genome sequencing among all three *SIRT6*^{-/-} monkey samples. All the possible off-target sites were identified by whole-genome sequencing regardless of the similarity to sgRNA sequences, as described in Methods.



Extended Data Fig. 2 | Characterization of *SIRT6*^{-/-} monkeys.

a, Western blots showing the absence of the SIRT6 protein and the alterations in H3K56ac levels in the kidney, lung, liver and skin of *SIRT6*^{-/-} monkeys. **b**, Immunohistochemical staining for H3K56ac in tissues from *SIRT6*^{-/-} and wild-type monkeys. The red dashed lines show

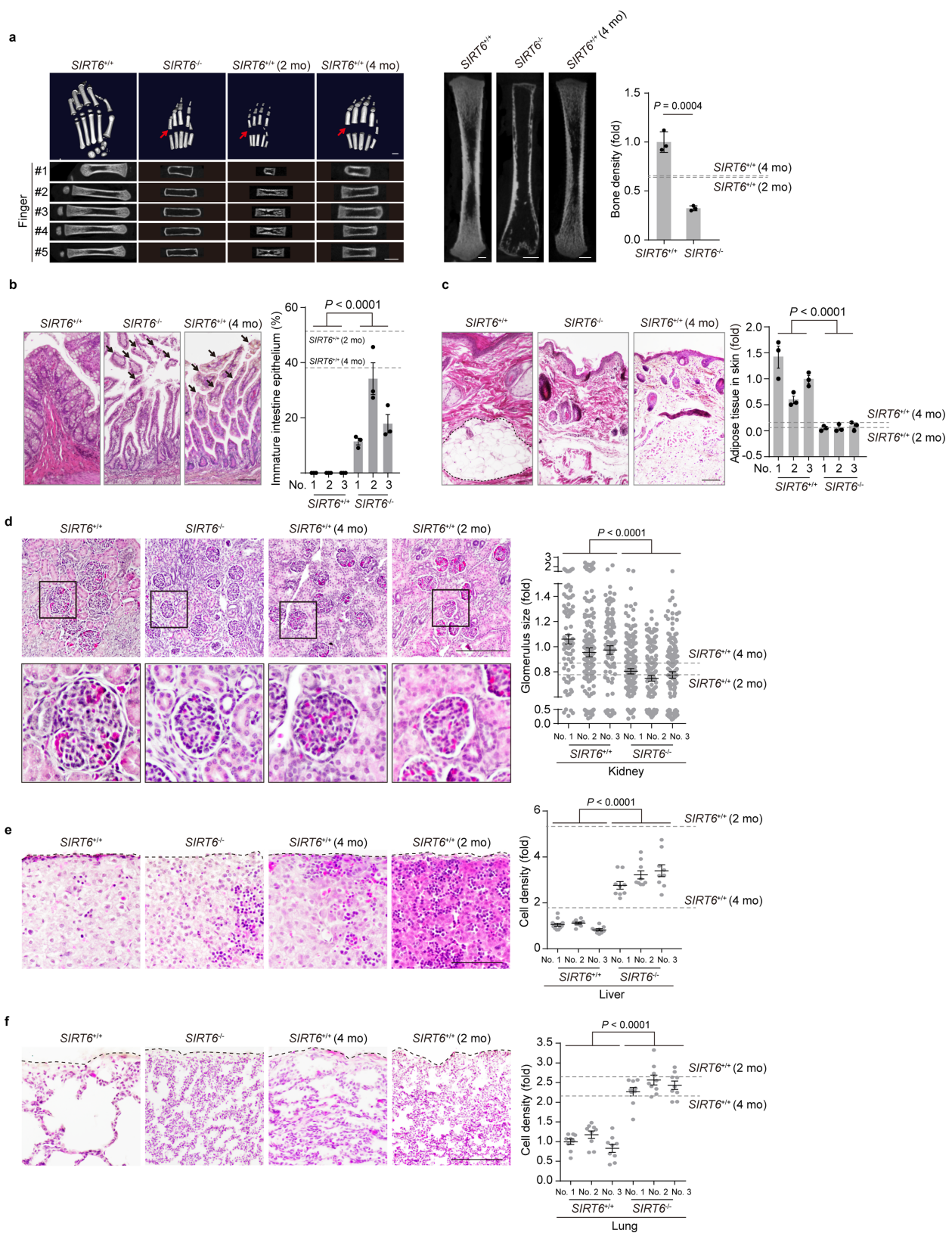
the upper boundaries of the cortices. Scale bar, 100 μ m. $n \geq 54$ glomeruli, kidney; $n = 18$ images, lung; $n = 18$ images, liver; $n = 18$ images, skin; $n = 9$ images, brain. Data are mean \pm s.e.m.; P values were determined by one-way ANOVA followed by Holm-Sidak's multiple comparisons test. For uncropped gels, see Supplementary Fig. 1.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | SIRT6 deficiency does not cause genomic and epigenomic instabilities in newborn monkeys. **a, b**, The relative telomere lengths of chromosomes from the monkey brain and liver were measured by qPCR. $n = 3$ monkeys. **c**, Whole-genome sequencing of copy number variations in brain samples from *SIRT6*^{+/+} and *SIRT6*^{-/-} monkeys. Each point represents a 500-kb genomic region of each chromosome. **d**, The distribution of different types of repetitive sequences in the monkey genome. Each type of repeat is marked in the same colour: LINE type 1 (LINE-1), SINEs, LTRs (including ERV1, ERV2, EVR3 and other types of LTR), rRNAs and other types of repeat (such as satellite, small nuclear RNA and 'other', as annotated in Repbase). The distributions of repeat reads were not significantly different in the wild-type and *SIRT6*^{-/-} monkeys, according to the two sided Wilcoxon signed-rank paired test. **e**, Whole-genome sequencing of SNVs using cortical samples from *SIRT6*^{+/+} and *SIRT6*^{-/-} monkeys. Sites with a heterozygosity percentage ranging between 0% and 30% were considered as SNV sites, whereas sites with a heterozygosity of >30% were considered single

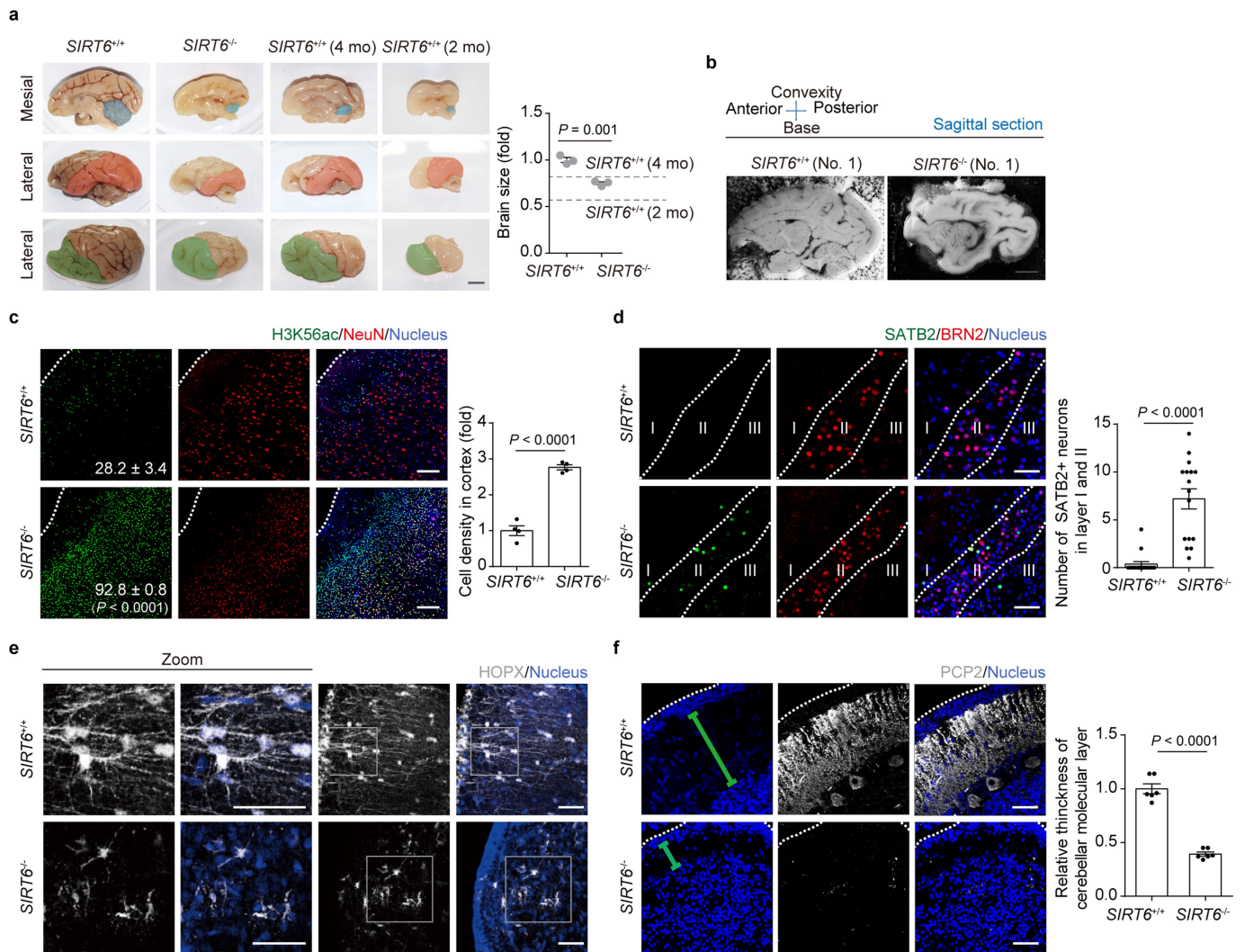
nucleotide polymorphisms. **f**, Schematic and morphology of primary mesenchymal stem cells isolated from monkey bone marrow. Scale bar, 50 μ m. **g**, Western blots showing the absence of the SIRT6 protein and the alterations in H3K56ac levels in primary mesenchymal stem cells from *SIRT6*^{-/-} monkeys. **h**, Karyotyping analysis of primary mesenchymal stem cells. **i**, Metaphase spread and fluorescence in situ hybridization showing telomeres in primary mesenchymal stem cells from wild-type and *SIRT6*^{-/-} monkeys (left). Quantification of the numbers of fluorescence in situ hybridization-labelled telomeres in each nucleus (right). $n \geq 21$ metaphases per monkey. Scale bar, 25 μ m. **j**, The relative telomere lengths of chromosomes from primary monkey mesenchymal stem cells were measured by qPCR. $n = 6$ qRT-PCR repeats for each monkey. **k**, Western blots of heterochromatin-related proteins in the cortex, muscle and liver of the wild-type and *SIRT6*^{-/-} monkeys. Data are mean \pm s.e.m., P values were determined by two-sided Student's t -test (**a, b**) or one-way ANOVA followed by Holm-Sidak's multiple comparisons test (**i, j**). For uncropped gels, see Supplementary Fig. 1.



Extended Data Fig. 4 | See next page for caption.

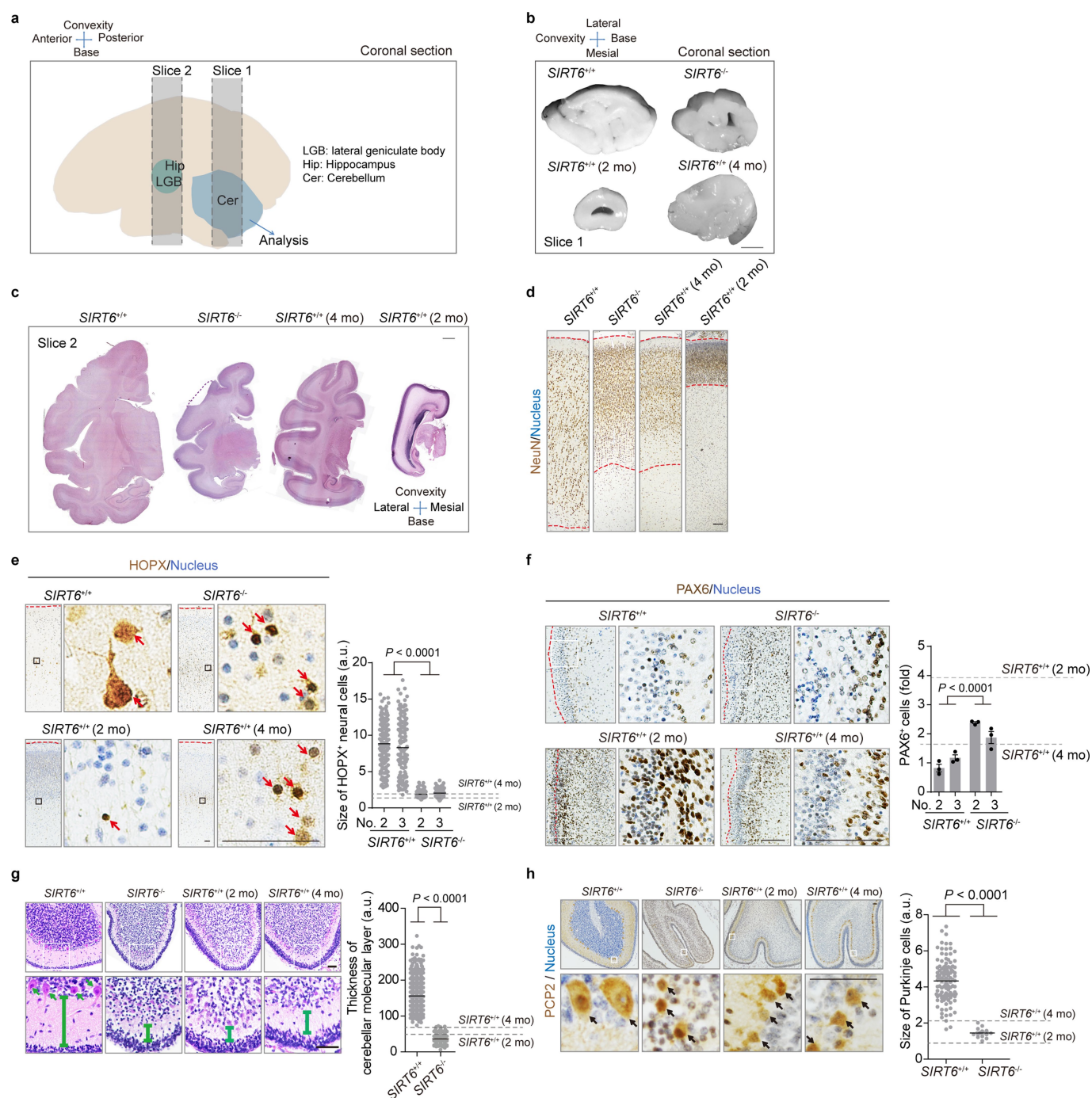
Extended Data Fig. 4 | SIRT6 deficiency resulted in a pan-tissue developmental delay. **a**, Micro-computed tomography of the hand (left), distal femoral trabecular (middle) bones and relative bone density (right) of wild-type and *SIRT6*^{-/-} newborn, and wild-type 2- and 4-month-old fetal, monkeys. Arrows point to the missing bone connections in the hands of the *SIRT6*^{-/-} monkeys and wild-type fetuses. Scale bars, 2.5 mm. *n* = 3 monkeys. **b**, Haematoxylin and eosin staining of the superficial intestinal epithelia from a wild-type newborn and 4-month-old fetus, and a *SIRT6*^{-/-} newborn monkey. Arrows point to regions of immature intestine epithelium in the tissues of the fetus and the *SIRT6*^{-/-} monkey. Scale bar, 75 μ m. *n* = 3 images (Vectra automated quantitative pathology imaging system) per monkey. **c**, Haematoxylin and eosin staining shows the

absence of subcutaneous fat in a newborn *SIRT6*^{-/-} monkey and wild-type 4-month-old fetus. In the tissue of the wild-type infant, the subcutaneous fat is circled with a dashed black line. Scale bar, 100 μ m. *n* = 3 images (Vectra automated quantitative pathology imaging system) per monkey. **d–f**, Haematoxylin and eosin staining of the indicated tissues in newborn *SIRT6*^{-/-} monkeys and 2-month-old fetal, 4-month-old fetal and newborn wild-type monkeys. **d**, Kidney, *n* \geq 80 glomeruli per monkey; **e**, liver, *n* \geq 9 images per monkey; **f**, lung, *n* = 9 images per monkey. In the bar graph, the grey dashed lines represent the average value of the wild-type fetuses. Scale bar, 100 μ m. Data are mean \pm s.e.m.; *P* values were determined by two-sided Student's *t*-test (**a**), one-way ANOVA followed by Holm–Sidak's multiple comparisons test (**b–f**).



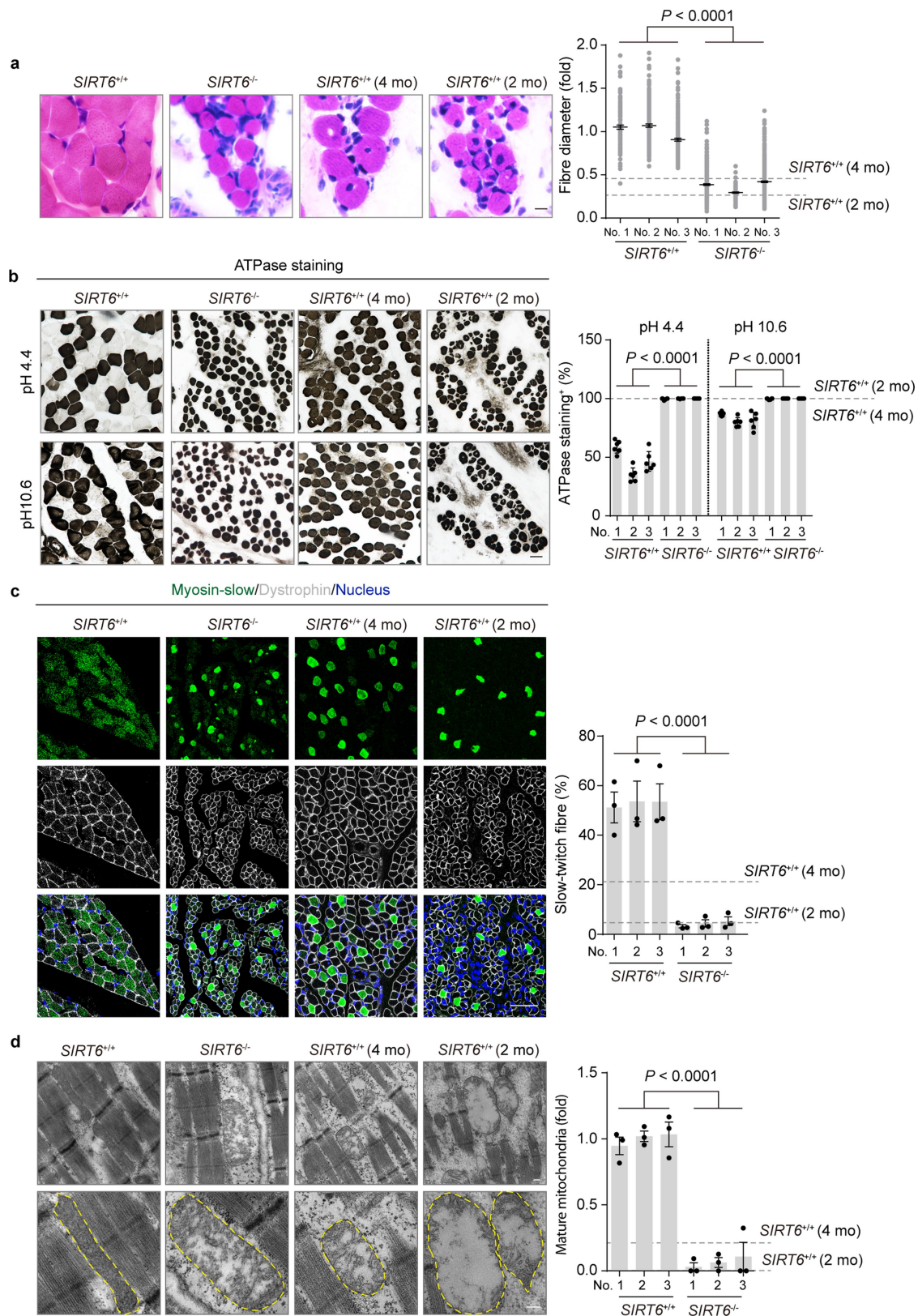
Extended Data Fig. 5 | SIRT6 deficiency resulted in a delay in brain development (sagittal sections). **a**, Representative photographs (left) and quantification (right) showing the sizes of the brains from wild-type newborn, wild-type 2-month-old and 4-month-old fetuses, and *SIRT6*^{-/-} monkeys. The cerebella are depicted in blue; the temporal and occipital lobes are depicted in red; and the frontal lobes are depicted in green. *n* = 3 monkeys. Scale bar, 0.5 cm. **b**, Representative photographs of sagittal sections from the indicated monkey. Scale bar, 0.5 cm. **c**, Left, images of immunofluorescence staining for H3K56ac and the neuronal nuclear antigen (NeuN) in the cortices of newborn wild-type and *SIRT6*^{-/-} monkeys. Right, cell density calculation. Numbers on the images represent percentages of H3K56ac-positive cells in the cortices. The white dashed lines identify the boundaries of the cerebral cortex. *n* = 4 images per monkey. **d**, Left, images of immunofluorescence staining

for SATB2 and BRN2 in the cortices of the newborn *SIRT6*^{-/-} and wild-type monkeys. Right, the proportion of SATB2⁺ cells in the cortical layers were quantified in brain sections. The white dashed lines identify the boundaries of the cortical layers. *n* = 16 images per monkey. **e**, Images of immunofluorescence staining for HOPX in *SIRT6*^{-/-} and wild-type cortical tissue sections. **f**, Images of immunostaining (left) for PCP2, a marker of cerebellar Purkinje neurons, and quantification of the thickness of the molecular layer of the cerebellum (right). PCP2⁺ cells were not detected in the cerebellum of *SIRT6*^{-/-} monkeys. The white dashed lines identify the boundaries of the cerebellum. Green lines indicate the thickness of the molecular layer of the cerebellum. Scale bar, 50 μ m (**c**–**e**). *n* = 6 images per monkey. Data are mean \pm s.e.m.; *P* values were determined by two-sided Student's *t*-test.



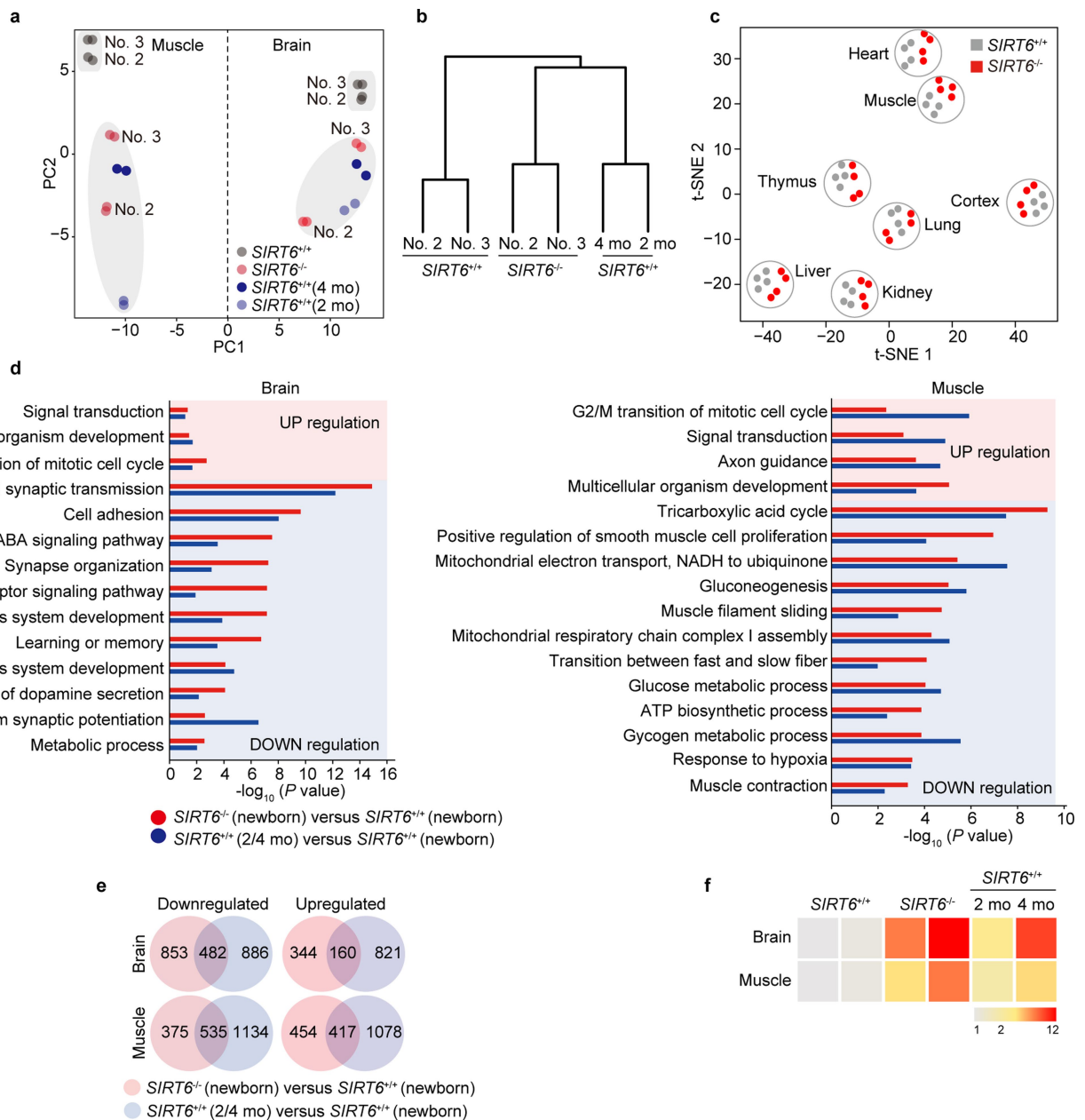
Extended Data Fig. 6 | SIRT6 deficiency resulted in a delay in brain development (coronal sections). **a**, Schematic of the brain regions analysed in our study. **b**, Representative photographs of coronal sections from the indicated monkeys. Scale bar, 0.5 cm. **c**, Haematoxylin and eosin staining in coronal sections from the brains of the indicated monkeys. Results shown are representative of three independent experiments. Scale bar, 0.25 cm. **d**, Images of immunostaining for the neuronal nuclear antigen (NeuN) in the cortices of *SIRT6*^{-/-} and wild-type newborn, and wild-type 2-month-old and 4-month-old fetal, monkeys. Results shown are representative of three independent experiments. Scale bar, 150 μ m. **e**, Left, images of immunostaining for HOPX in *SIRT6*^{-/-} and wild-type cortical tissue sections. Right, bar graph showing the diameter of the HOPX⁺ cells. The red arrows identify HOPX⁺ cells. *n* = 141. Scale bars, 100 μ m. **f**, Left, images of PAX6 immunostaining in the hippocampi of *SIRT6*^{-/-} and wild-type newborn monkeys, and wild-type fetuses. Right, bar graph showing the relative percentages of PAX6⁺ cells in brain sections. *n* = 3 slices per monkey. Scale bar, 100 μ m. **g**, Left, haematoxylin and eosin staining of the cerebellum from the wild-type and *SIRT6*^{-/-}

newborn, and wild-type 2-month-old and 4-month-old fetal, monkeys. Right, the relative thicknesses of the molecular layers. The bottom panels show higher magnification images of the boxed areas in the corresponding top panels. Green lines indicate the thickness of the molecular layer of the cerebellum. Green arrows identify Purkinje neurons. *n* \geq 156 images per genotype. Image was taken at about every 4 Purkinje neurons in wild-type newborn monkeys. Scale bar, 50 μ m. **h**, Images of immunostaining (left) and quantification of the size (right) of PCP2⁺ cerebellar Purkinje neurons. PCP2⁺ cells were very scarce (only 14 were found) in the tissue(s) from the *SIRT6*^{-/-} monkeys, *n* = 114 PCP2⁺ cells in wild-type monkeys. Arrows identify PCP2⁺ neurons. Scale bar, 50 μ m. Red dashed lines identify the boundaries of the cerebral cortex. Grey dashed lines represent the average values for the 2-month-old and 4-month-old fetuses. White box indicates enlarged area. a.u., arbitrary units. Horizontal lines show the average values for each group (e, g, h). Data are mean \pm s.e.m. (f); *P* values were determined by two-sided Student's *t*-test or one-way ANOVA followed by Holm-Sidak's multiple comparisons test.



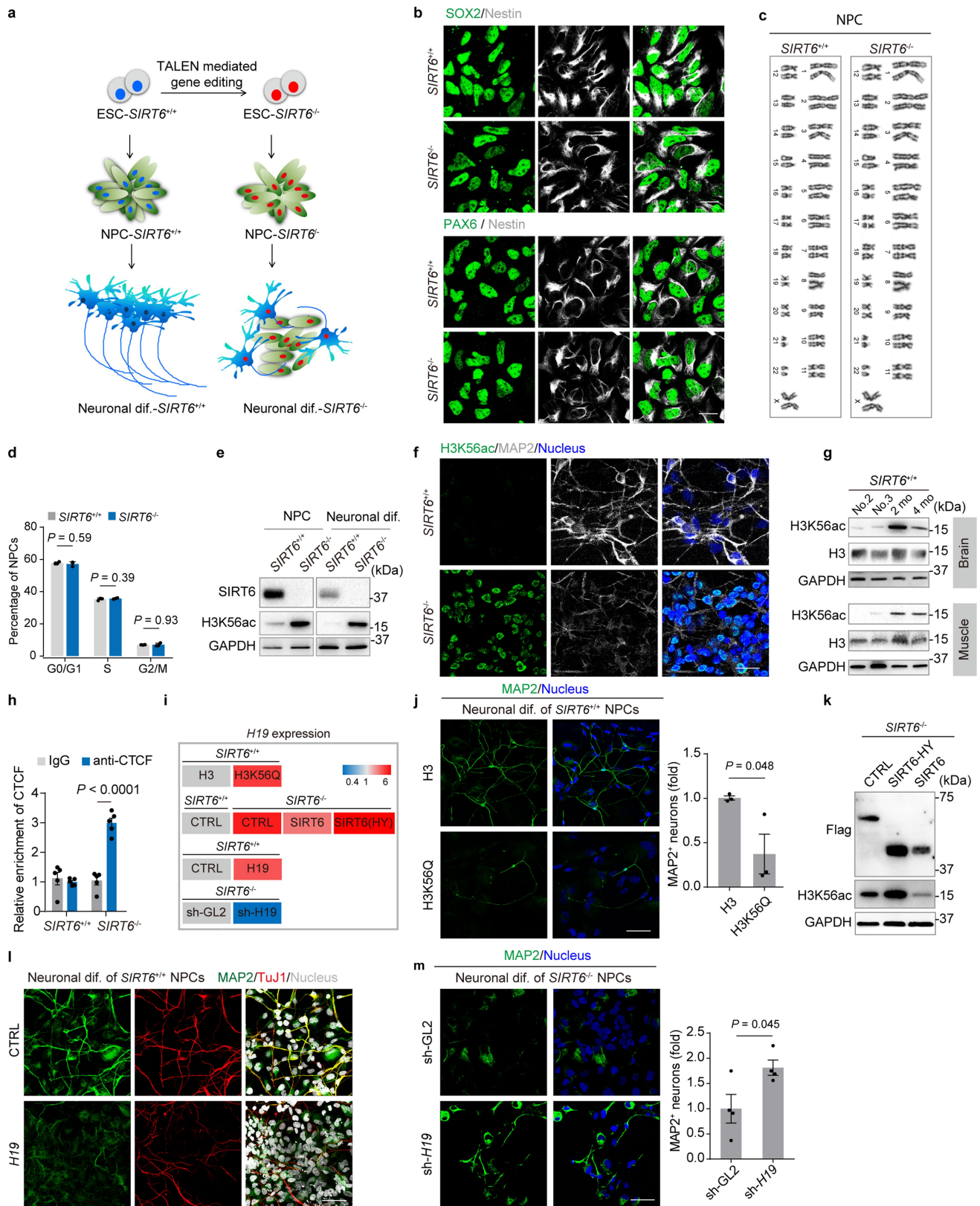
Extended Data Fig. 7 | SIRT6 deficiency resulted in a delay in muscle development. **a**, Representative images of haematoxylin and eosin staining in the gastrocnemius and soleus muscles from the indicated monkeys, and quantification of the fibre diameter ($n \geq 60$ fibres). Scale bar, 10 μm . **b**, Images (left) and quantification (right) of myosin ATPase staining at pH 4.4 and 10.6 in cross-sections of the gastrocnemius and soleus muscle tissue. $n = 6$ slices per monkey. Scale bar, 20 μm . **c**, Left, images of immunofluorescence staining for the slow-twitch marker myosin-slow and sarcolemma-bound protein dystrophin in the gastrocnemius and soleus

muscle. Scale bar, 50 μm . Right, distributions of the types of muscle fibres were calculated as a percentage of myosin-slow⁺ fibres. $n = 3$ slices per monkey. **d**, Electron microscopy images of mitochondrial morphology showing dense mitochondrial cristae predominantly in infant wild-type monkeys. Each yellow dashed circle identifies a single mitochondrion. Grey dashed lines represent the average values for the 2-month-old and 4-month-old fetuses. Scale bar, 100 nm. $n = 3$ slices per monkey, ≥ 30 mitochondria per slices. Data are mean \pm s.e.m.; P values were determined by one-way ANOVA followed by Holm–Sidak’s multiple comparisons test.



Extended Data Fig. 8 | *SIRT6*^{-/-} monkeys exhibit embryonic transcriptome features. **a**, Principal component analyses of brain and muscle tissues from *SIRT6*^{+/+} newborn and fetal monkeys and *SIRT6*^{-/-} monkeys, based on development-related genes as described in Methods. $n = 2$ independent experiments per tissue per monkey. **b**, Hierarchical clustering analysis based upon global DNA methylation of brain tissues from *SIRT6*^{+/+} newborn and fetal monkeys and *SIRT6*^{-/-} monkeys, as described in Methods. **c**, An unsupervised t-SNE cluster analysis of the *SIRT6*^{-/-} (red) and wild-type (grey) samples. $n = 2$ independent experiments per tissue per monkey. **d**, The Gene Ontology terms for genes in *SIRT6*^{-/-} newborn monkeys and wild-type fetuses that were differentially expressed in the same manner as in the wild-type infants.

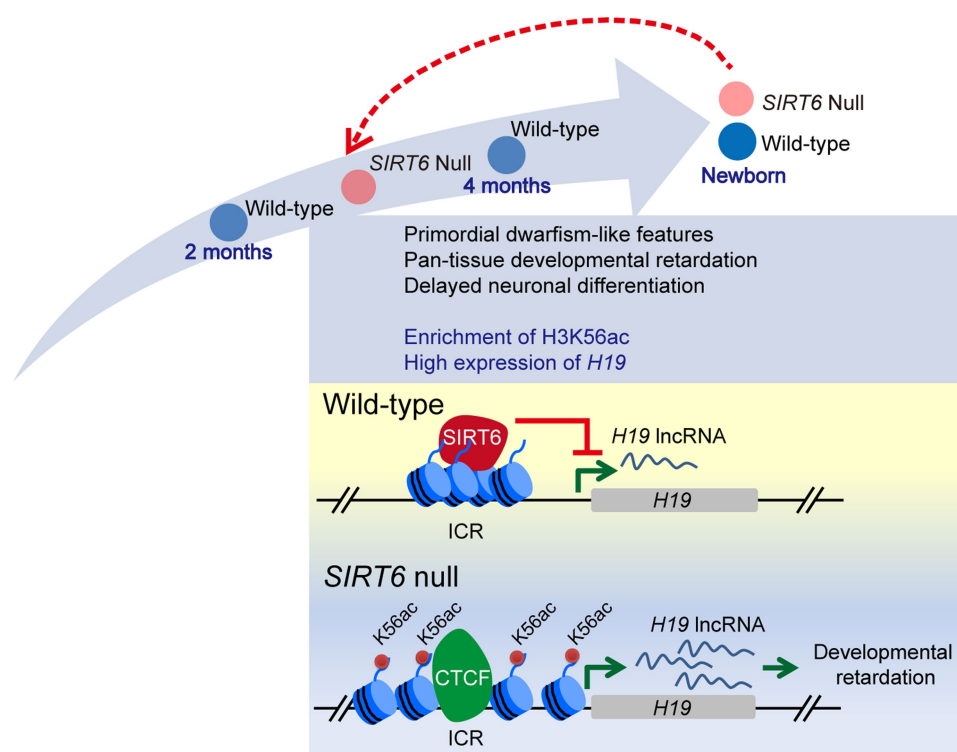
e, Venn diagrams showing the overlap among differentially upregulated and downregulated genes (fold change (*SIRT6*^{-/-}/*SIRT6*^{+/+}) > 2 or < 0.5, false-discovery-rate-adjusted $P < 0.05$) in brains and muscles of the *SIRT6*^{-/-} and wild-type infants, and the 2-month-old and 4-month-old wild-type fetuses and the wild-type infants. $n = 2$ independent experiments per tissue per monkey. **f**, Levels of *H19* in muscle and brain tissues from the wild-type newborn (no. 2 and no. 3) and fetal monkeys and *SIRT6*^{-/-} monkeys, as measured by qRT-PCR. For **d**, **e**, the average transcript levels in the tissues from *SIRT6*^{+/+} monkeys no. 2 and no. 3 were normalized to 1, and the relative expression level of each gene was coloured as shown in the bottom panel.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Deacetylation of H3K56 by SIRT6 regulates *H19* expression. **a**, Schematic showing the method used to generate *SIRT6*^{+/+} and *SIRT6*^{-/-} NPCs, which were differentiated into post-mitotic neurons. **b**, Images of immunofluorescence staining of *SIRT6*^{+/+} and *SIRT6*^{-/-} NPCs. Scale bar, 20 μ m. **c**, Karyotyping analysis of NPCs. **d**, Measurements of the percentages of early passage NPCs at each cell cycle stage, $n = 3$ wells per condition. **e**, Western blots of the indicated proteins in the *SIRT6*^{+/+} and *SIRT6*^{-/-} NPCs and in their neuronal differentiation products (neuronal dif.). **f**, Immunofluorescence staining for H3K56ac levels in *SIRT6*^{+/+} and *SIRT6*^{-/-} neuronal differentiation products. Scale bar, 50 μ m. **g**, Western blots showing H3K56ac levels in protein extracts from the cortices and muscle from the infant and 2-month-old and 4-month-old fetal *SIRT6*^{+/+} monkeys, and from *SIRT6*^{-/-} newborn monkeys. **h**, ChIP-qPCR showing the enrichment of CTCF at the imprinting control region of *H19* in the genomes of differentiated neuronal cells. $n = 5$ wells per condition. **i**, qPCR assessment of the *H19*

levels in neuronal derivatives transduced with the indicated lentiviral vector. $n = 4$ wells per condition. **j**, Images of immunofluorescence staining (left) and statistical analyses (right, bar graph) of the neuronal cells that overexpress H3(K56Q). $n = 3$ wells per condition. Scale bar, 50 μ m. **k**, Western blots of levels of H3K56ac in protein extracts from *SIRT6*^{-/-} NPCs that overexpress luciferase (control), inactive SIRT6 (SIRT6-HY) and wild-type SIRT6. **l**, Images of immunostaining for MAP2 and TuJ1 in the neuronal derivatives of wild-type NPCs that overexpress *H19* (see Fig. 4e). Scale bar, 50 μ m. **m**, Images of immunofluorescence staining (left) and statistical analyses (right, bar graph) of the neuronal cells with *H19* shRNA vector ($n = 4$ wells per condition). Scale bar, 50 μ m. Data are mean \pm s.e.m., P values were determined by two-sided Student's t -test (**i**, **j** and **m**) or one unpaired t -test followed by multiple comparisons using the Holm-Sidak method (**d**, **i**). For uncropped gels, see Supplementary Fig. 1.



Extended Data Fig. 10 | A model of the mechanism by which SIRT6 regulates prenatal development in monkeys. Top, SIRT6-null monkeys die soon after birth and exhibit retarded development that resembles that of 2- to 4-month-old fetuses. Notably, SIRT6 deficiency delays neuronal development, which was recapitulated in an in vitro study examining the differentiation of wild-type and *SIRT6*^{-/-} human NPCs. Bottom, during

neuronal differentiation SIRT6 is recruited to the imprinting control region of *H19*, where it deacetylates H3K56ac and thereby prevents CTCF-mediated transcription of the developmental repressor long non-coding RNA *H19*. SIRT6 depletion promotes the expression of *H19* and arrests neuronal differentiation.

Crystal structure of the Frizzled 4 receptor in a ligand-free state

Shifan Yang¹, Yiran Wu^{1,11}, Ting-Hai Xu^{2,11}, Parker W. de Waal^{2,11}, Yuanzheng He³, Mengchen Pu¹, Yuxiang Chen^{1,4,5,6}, Zachary J. DeBruine², Bingjie Zhang¹, Saheem A. Zaidi⁷, Petr Popov^{7,8}, Yu Guo^{1,4,5,6}, Gye Won Han⁷, Yang Lu³, Kelly Suino-Powell², Shaowei Dong^{1,4,5,6}, Kaleeckal G. Harikumar⁹, Laurence J. Miller⁹, Vsevolod Katritch^{7,8}, H. Eric Xu^{2,10}, Wenqing Shui^{1,4}, Raymond C. Stevens^{1,4}, Karsten Melcher², Suwen Zhao^{1,4} & Fei Xu^{1,4*}

Frizzled receptors (FZDs) are class-F G-protein-coupled receptors (GPCRs) that function in Wnt signalling and are essential for developing and adult organisms^{1,2}. As central mediators in this complex signalling pathway, FZDs serve as gatekeeping proteins both for drug intervention and for the development of probes in basic and in therapeutic research. Here we present an atomic-resolution structure of the human Frizzled 4 receptor (FZD4) transmembrane domain in the absence of a bound ligand. The structure reveals an unusual transmembrane architecture in which helix VI is short and tightly packed, and is distinct from all other GPCR structures reported so far. Within this unique transmembrane fold is an extremely narrow and highly hydrophilic pocket that is not amenable to the binding of traditional GPCR ligands. We show that such a pocket is conserved across all FZDs, which may explain the long-standing difficulties in the development of ligands for these receptors. Molecular dynamics simulations on the microsecond timescale and mutational analysis uncovered two coupled, dynamic kinks located at helix VII that are involved in FZD4 activation. The stability of the structure in its ligand-free form, an unfavourable pocket for ligand binding and the two unusual kinks on helix VII suggest that FZDs may have evolved a novel ligand-recognition and activation mechanism that is distinct from that of other GPCRs.

FZDs have been grouped as class-F GPCRs on the basis of their sequence homology with other GPCR families. Although the main intracellular FZD-binding protein and Wnt signal-transducer is Dishevelled, several FZD receptors also have the ability to couple to G proteins and induce G-protein nucleotide exchange, which defines them as GPCRs³. However, the mechanistic details of FZD signalling through G proteins are poorly understood, and it is not clear whether all FZDs can couple to G proteins. FZDs are involved in the regulation of many biological processes during embryonic development and tissue homeostasis⁴. As the major cell-surface receptor for the Wnt pathway, ten FZDs have central roles in Wnt signalling, and their aberrance is linked to numerous diseases^{5–7}. Consequently, the structural and functional characterization of FZD and the exploration of FZD-targeted therapeutics have received considerable attention in recent years. The elucidation of the structures of an extracellular soluble region in FZDs—the cysteine-rich domain (CRD)^{8,9}—and its complex with Wnt protein^{10–12} led to the discovery of Wnt surrogates, or antibody fragments, that target this region^{13,14}. However, ligands that target the transmembrane domain (TMD) of FZDs—the traditional pocket in which many GPCR-targeting drugs act—are scarce, and the dynamics and signalling of these receptors are still largely unknown. To provide guidance on the development of FZD ligands and to understand the

activation and signalling of FZDs, we solved the TMD structure of the human FZD4 receptor (hereafter referred to as Δ CRD-FZD4)—which shares 37–61% sequence homology with the other nine FZDs (Extended Data Fig. 1)—at a resolution of 2.4 Å in a ligand-free state (Extended Data Table 1a).

The construct of Δ CRD-FZD4 contains residues 178–517, with eight residues (420–427) of the third intracellular loop (ICL3) replaced by rubredoxin to facilitate crystallization (see Methods and Extended Data Fig. 2). The overall structure of Δ CRD-FZD4 reveals a canonical GPCR fold with seven transmembrane helices (7TM; helices I–VII) and a short helix 8 (H8) containing the conserved signalling-related KXXXXW motif, which is essential for the recruitment of the intracellular signal transducer Dishevelled¹⁵, packed parallel to the membrane bilayer (Fig. 1a). We crystallized Δ CRD-FZD4 in its ligand-free state, and to our knowledge this is the first apo structure of a ligand-regulated GPCR (Fig. 1b).

Compared to the smoothened receptor (SMO), which belongs to the class-F receptors and is an important target for antitumour therapy^{16–18}, the 7TM bundle fold of FZD4 is similar, with a C_{α} root mean square deviation (r.m.s.d.) of 1.2 Å. Despite the high structural conservation of the TMDs, the helix-VI extension in FZD4 is much shorter than that in SMO. Additionally, the intracellular end of helix V moves outwards from the helical bundle by about 13° (Fig. 1c). To assess the stability of helices V and VI and their role in receptor activation, we ran three independent three-microsecond molecular dynamics simulations based on our structure (Extended Data Fig. 3a). We observed that helix VI of FZD4 remained closely packed within the TMD helical bundle of the receptor throughout simulation. This locked conformation of helix VI is primarily mediated by substantial hydrophobic interactions and two unique backbone hydrogen bonds (W320^{3,43f} and W327^{3,50f}; superscripts denote Ballesteros–Weinstein numbering for GPCRs) that connect helix III to helix VI (Extended Data Fig. 3b). Unlike most of the other class-A GPCRs and SMO¹⁹, in which helix VI moves outwards upon activation, helix VI in FZD4 remained relatively stable in an inward orientation (Extended Data Fig. 3b).

Among the ten FZDs, FZD4 is the only receptor that can be activated by binding of the protein Norrin. It is involved in retinal angiogenesis and in maintaining the integrity of the blood–retinal barrier, and mutations of FZD4 are found in familial exudative vitreoretinopathy^{20,21}. To gain insight into this structure–function relationship, we mutated some disease- and signalling-related ‘hotspot’ residues²² and assessed their function using the TOPflash reporter assay. We found that these mutations were in key positions of our Δ CRD-FZD4 structure and led to aberrant downstream signalling (Extended Data Fig. 4a). We also observed that family-conserved amino acids (Y250^{2,39f} and W494^{7,55f})

¹Human Institute, ShanghaiTech University, Shanghai, China. ²Center for Cancer and Cell Biology, Innovation and Integration Program, Van Andel Research Institute, Grand Rapids, MI, USA. ³HIT Center for Life Sciences, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China. ⁴School of Life Science and Technology, ShanghaiTech University, Shanghai, China. ⁵Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. ⁶University of Chinese Academy of Sciences, Beijing, China. ⁷Departments of Biological Sciences and Chemistry, Bridge Institute, University of Southern California, Los Angeles, CA, USA. ⁸Moscow Institute of Physics and Technology, Dolgoprudny, Russia. ⁹Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, AZ, USA. ¹⁰Key Laboratory of Receptor Research, VARI-SIMM Center, Center for Structure and Function of Drug Targets, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. ¹¹These authors contributed equally: Yiran Wu, Ting-Hai Xu, Parker W. de Waal. *e-mail: xufe@shanghaitech.edu.cn

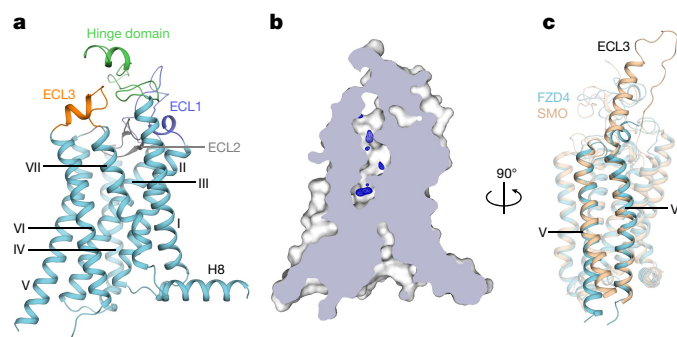


Fig. 1 | Overall structure of Δ CRD-FZD4 and the comparison with SMO. **a**, Side view of Δ CRD-FZD4 with the hinge domain, ECL1, ECL2 and ECL3 domains coloured in green, blue, grey, and orange, respectively. **b**, The electron density in the pocket represented as an $|F_o| - |F_c|$ maps contoured at 3.0σ in dark blue, which indicates that our structure is in a truly apo state (residual densities represent water molecules and an unknown atom or ion density). **c**, Comparison of Δ CRD-FZD4 (cyan) with Δ CRD-SMO (gold, truncated from RCSB Protein Data Bank code (PDB): 5L7D).

adopted different conformations in Δ CRD-FZD4 when compared with SMO (Extended Data Fig. 4b). These residues have been reported to have important roles in downstream signalling and in maintaining structural integrity, which indicates that these key residues have evolved distinct conformations in different proteins to correspond to their respective functions²³. Our mutational analysis confirmed that the mutations Y250F and W494L led to reduced signalling activity (Extended Data Fig. 4a).

Despite the lack of the CRD, the arrangement of the extracellular side (hinge domain and extracellular loops (ECLs)) of this Δ CRD-FZD4 structure reveals some interesting features of the domain interaction and sheds light on the connection to the CRD. Therefore, we compared

the domain organization of FZD4 with that of other multi-domain GPCRs. For the two class-F receptors, FZD4 and SMO, individual fragments of the extracellular side (hinge domain, ECL1, ECL2 and ECL3) pack together to form a compact structure (Fig. 2a, b). This unique tertiary structure in turn stabilizes the 7TM bundle with a continuous interface of $2,625 \text{ \AA}^2$. In FZD4, the ECL2 β -hairpin acts as a plug, embedded deeply in the receptor, and occupies an abundance of space in the 7TM bundle cavity. ECL1 and ECL3 sandwich and stabilize the hinge domain through polar and nonpolar interactions (Fig. 2a). By contrast, although the ECL2 β -hairpin is located at a similar position in SMO, the hinge domain stacks on top of ECL1 and leans against ECL3, adopting an elongated conformation (Fig. 2b). This difference in organization of the hinge domain between FZD4 and SMO may result in a different arrangement of or different dynamics towards their respective CRDs. To illustrate this, we built a tentative full-length model of FZD4 (Extended Data Fig. 5a), which shows that the much shorter ECL3 of FZD4 markedly reduces the interface area between the CRD and TMD surfaces compared to SMO. Such a diminished interface results in a less stable orientation of the CRD, which we have confirmed by molecular dynamics simulations: substantial swinging of the domain is observed on a 1.5-microsecond scale (Extended Data Fig. 5b).

Distinct from class-F receptors, the extracellular regions (hinge domain and extracellular loops) in class-B and class-C receptors are less compact. In the glucagon receptor (GCGR)²⁴, the glucagon-like peptide 1 receptor (GLP-1R)²⁵ and the metabotropic glutamate-1 receptor (mGlu-1R)²⁶, the hinge domain interacts tightly with ECL1, the partially disordered ECL2 moves away from the binding cavity, and ECL3 is in an isolated position with no contact with other members from the extracellular side (Fig. 2c–e). Not surprisingly, such loose stacking leaves a spacious binding pocket for cognate ligands and supports an ‘open–closed’ conformational change of the extracellular domain in the activation of class-B GPCRs.

We next compared the arrangement of the intracellular side of FZD4 with that of SMO and representative class-A receptors (Extended

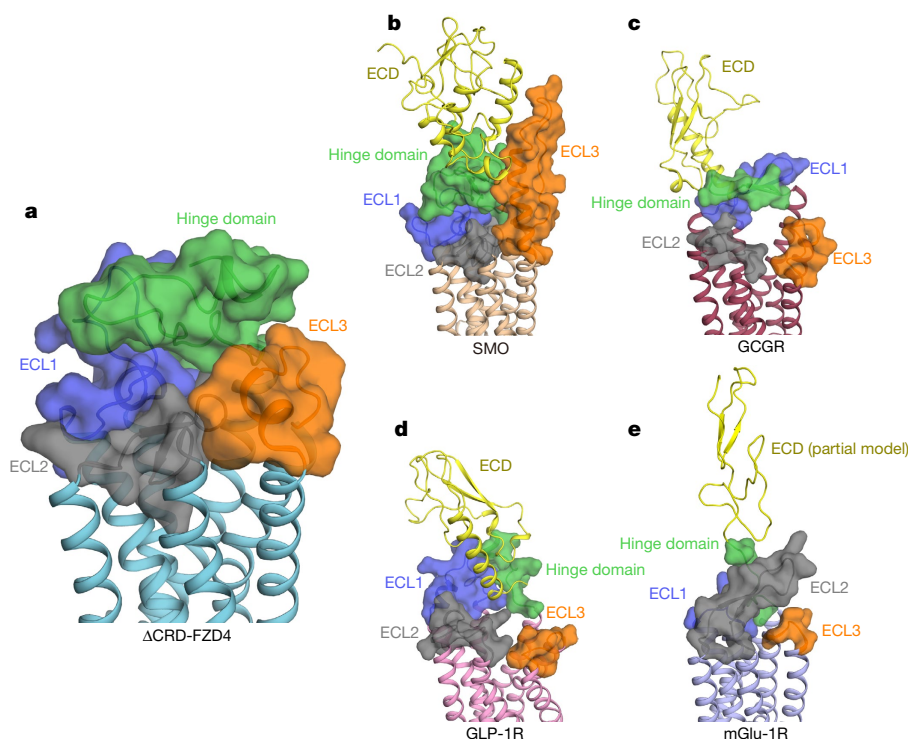


Fig. 2 | The arrangement of the extracellular side of FZD4 and other GPCRs. **a**, **b**, In FZD4 (**a**) and SMO (**b**; PDB: 5L7D), ECL2 acts as a plug and the individual fragments of the extracellular region (hinge domain, ECLs) form a compact structure. **c**–**e**, In GCGR (PDB: 5XF1), GLP-1R (PDB: 5NX2) and mGlu-1R (PDB: 4OR2), the organization of the

extracellular region (hinge domain and ECLs) is less compact so either an antibody or a ligand is required for stabilization of the 7TM and/or the extracellular domain (ECD). ECDs for SMO, GCGR and GLP-1R are derived from full-length structures, and for mGlu-1R a partial model (PDB: 2E4W) is presented.

Data Fig. 3c). It was noted that FZD4 is flatter on the intracellular side compared to SMO and some class-A GPCRs, and that no cavity exists among helices III, VI or VII; such a cavity has been identified as an allosteric ligand-binding site in some class-A GPCRs²⁷. The family-conserved Arg^{3,50} in class-A GPCRs, which has previously been reported to be a G-protein-binding site²⁸, is also missing in FZD4 and SMO. The various different arrangements in FZD4 suggest that this receptor has a different mechanism for the recognition of allosteric ligands and downstream signalling molecules compared with other GPCRs.

Despite being a central intervention point of the Wnt–FZDs pathway, the discovery of a ligand that directly targets the traditional GPCR transmembrane pocket of FZDs has been challenging. Nonetheless, several drugs that target the SMO transmembrane pocket have been approved by the US Food and Drug Administration for the treatment of cancer. Compared to that of SMO, the transmembrane pocket of FZD4 is more constricted and cannot accommodate SMO ligands. When we superimposed SMO ligands, extensive clashes were found between the ligands and the FZD4 pocket (Fig. 3a and Extended Data Fig. 6a); this result was further confirmed by affinity mass spectrometry to assess the receptor–ligand interaction (Extended Data Table 1b). To understand these extensive clashes, we calculated the volume of the transmembrane pocket in FZD4, SMO and in all other GPCRs for which structures have been solved. The volume of the FZD4 pocket (882 Å³) is larger than that of the SMO pocket (763 Å³), and is in the middle of the range for all GPCRs with solved structures. However, there are two extremely narrow points—caused by the presence of bulky side chains—in which the area of the channel of the FZD4 pocket is reduced to 4.8 Å² and 8.0 Å² (Fig. 3b, c). Further molecular dynamics simulation analysis revealed that the side chains pointing to the pocket could only fluctuate over a small range, and that the cross-sectional area of the pocket does not change considerably (Fig. 3d and Extended Data Fig. 6b). When we mutated these residues and tested the effects by cell-based TOPflash reporter assay, we found that none of the mutations markedly reduced FZD4 activation either in the absence or in the presence of endogenous agonist (Fig. 3e and Extended Data Fig. 6c), which is consistent with our hypothesis that agonist binding to the pocket is not required for canonical signalling.

Besides being markedly different from its analogues in terms of shape, the pocket of FZD4 also contains a cluster of polar residues pointing to the cavity, which makes it the most hydrophilic pocket of all GPCRs with known structure (Fig. 3d and Extended Data Fig. 6d). Conservation analysis of homology models for the other nine FZDs based on our structure (Extended Data Fig. 7) indicated that all FZDs share high structural similarity and sequence homology at transmembrane helices, whereas the extracellular and intercellular regions are poorly conserved (Extended Data Fig. 7a). The aforementioned pocket-shape-related and polar residues form a conserved hydrophilic pocket in all FZDs that may substantially reduce the binding affinity of a ligand (Extended Data Fig. 7b–d). This may partially explain why it has been a challenge to develop good small-molecule ligands for FZDs in recent years. Other challenges arise from the redundancy of FZD in the genome and the high degree of homology among the ten FZDs, which might prohibit the identification of ligands as specific small-molecule inhibitors for each individual FZD.

To understand the activation and signalling of FZD4, we carried out a series of molecular dynamics simulations on a three-microsecond timescale, in addition to mutagenesis studies. Molecular dynamics results indicate a concerted movement at the proposed Dishevelled-binding site²³, where the receptor cycled between a ‘closed’ and a ‘bent’ conformation (Fig. 4a and Extended Data Fig. 8a). When cycling between these two states, the hydrogen bond between W352^{4,50f} and H348^{4,46f} breaks, which enables H348^{4,46f} to swing by around 5 Å. Helix IV adopts a bent conformation enabling Y250^{2,39f}, which is sandwiched by R253^{2,42f} and E341^{4,39f} in the crystal structure (closed conformation), to swing by approximately 5 Å towards the newly exposed backbone carbonyl of S344^{4,42f} (Fig. 4a). This opens a new

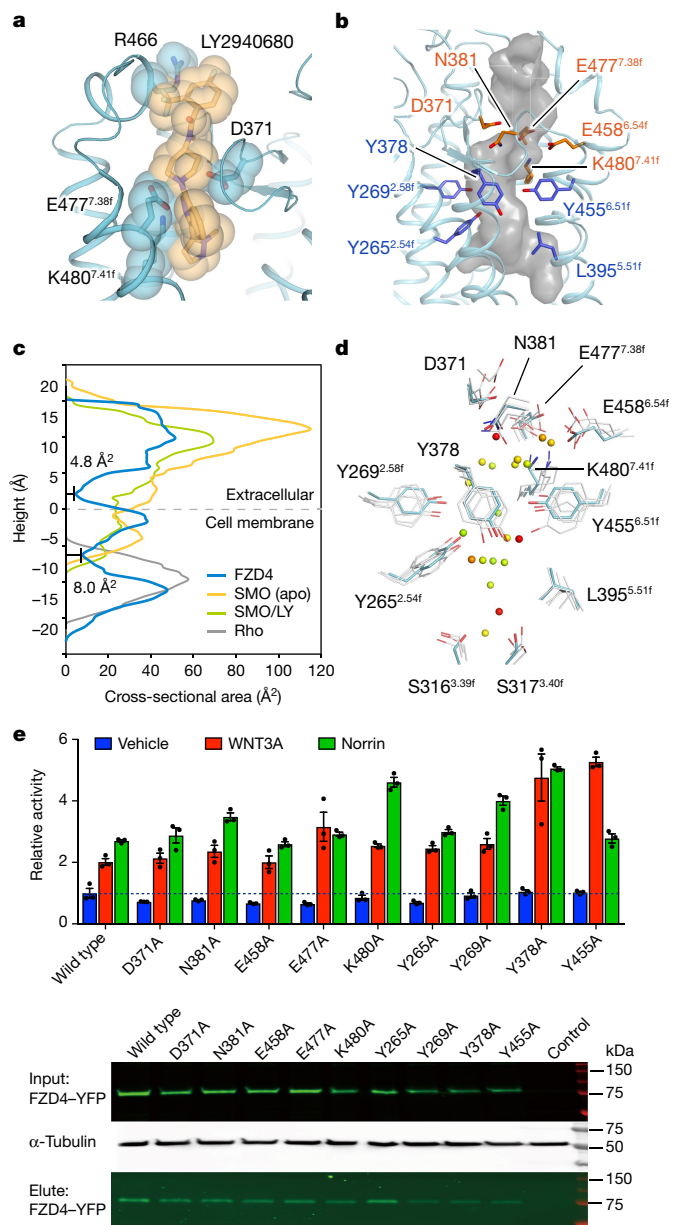


Fig. 3 | Transmembrane ligand pocket of FZD4. **a**, Superimposition of the SMO ligand LY2940680 (gold) in the pocket of FZD4 (cyan). Note the extensive clashes between LY2940680 and FZD4. **b**, Two narrow positions in the pocket of FZD4. The amino acids clamping position I (upper position) are shown in orange and those clamping position II (lower position) are shown in blue. **c**, The cross-sectional area of the ligand pockets of FZD4 (blue), the apo form of SMO (orange), SMO bound to LY2940680 (SMO/LY; green) and rhodopsin (Rho; grey). Note that there are two narrow positions in the pocket of FZD4. **d**, Molecular dynamics simulations of water molecules and key pocket-residue rotamers in the FZD4 pocket. Pocket-residue side chains in the crystal structure and in molecular dynamics simulations are coloured in cyan and grey, respectively. Simulated water molecules are represented as small spheres and coloured according to frequency: darker colours indicate higher frequencies. **e**, Mutational analysis of residues pointing towards the pocket indicated that agonist binding to the pocket is not required for signalling. The TOPflash reporter gene assay (top) was normalized to the activity of unstimulated wild-type FZD4 and each data point represents the mean \pm s.e.m., repeated in triplicate. The cell-surface expression (bottom) for each mutant was determined by isolation of cell-surface-biotinylated proteins on avidin resin (see Methods). Cell-surface biotinylation was cross-validated by fluorescence microscopy analysis with similar results. Corresponding cell localization and total expression data are shown in Extended Data Fig. 6c.

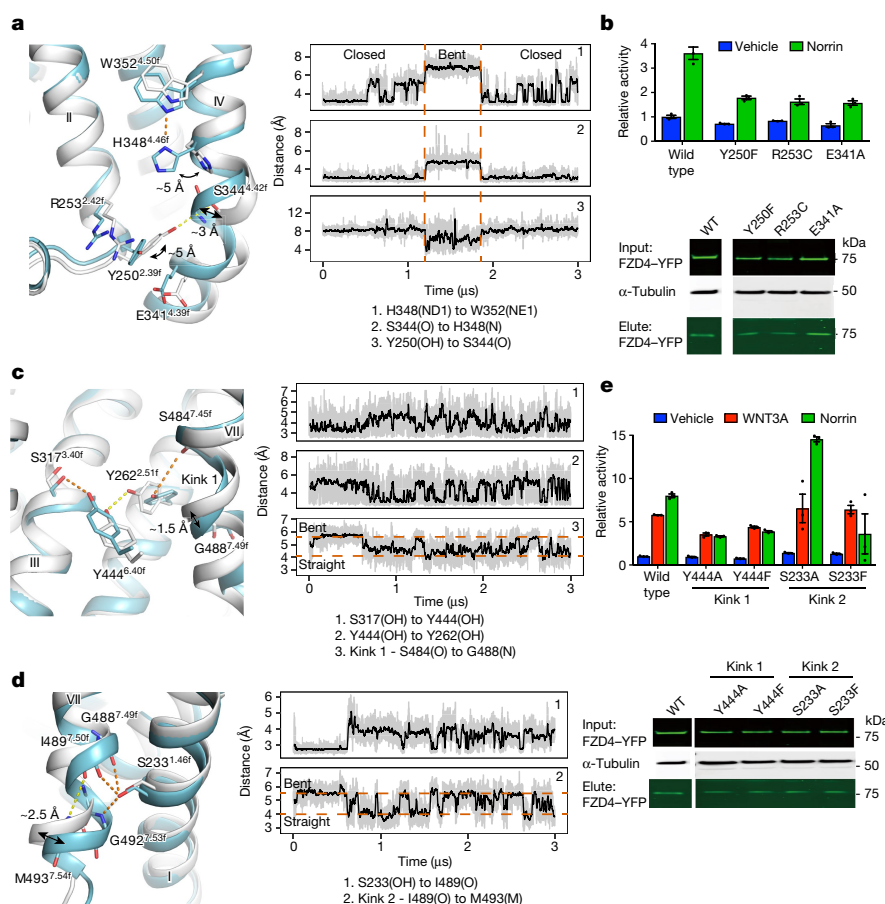


Fig. 4 | Activation (kinks) and Dishevelled-mediated signalling analysis of FZD4. **a**, Snapshots and distance traces of closed (cyan) and bent (grey) conformations observed during simulation at the Dishevelled-binding site. **b**, Activity assay (TOPflash reporter assay) and cell-surface expression data for mutations of the Y250 network residues. Each data point represents the mean \pm s.e.m., repeated in triplicate. Cell-surface expression data are shown underneath the bar chart. Cell-surface biotinylation was cross-validated by fluorescence microscopy analysis with similar results. **c**, **d**, Snapshots and distance traces of bent (cyan) and straight (white)

conformations for kink 1 (**c**) and kink 2 (**d**) in helix VII. **e**, Activity assay (TOPflash reporter assay, top) and cell-surface expression data (bottom) for mutations in kink1 and kink2. Each data point represents the mean \pm s.e.m., repeated in triplicate. Cell-surface biotinylation was cross-validated by fluorescence microscopy analysis with similar results. Corresponding cell localization and total expression data are shown in Extended Data Figs. 8 and 9. Distance traces from molecular dynamics data are plotted as a moving average over a 10-ns window.

polar cavity that is flooded with additional water molecules (Extended Data Fig. 8a). Furthermore, mutations of any of the residues in the R253^{2.42f}–Y250^{2.39f}–E341^{4.39f} sandwich caused Norrin-induced activation defects (Fig. 4b and Extended Data Figs. 4b, 8b). Our results suggest that this region can undergo a large conformational change in which the cytoplasmic portion of helix IV swings away from the receptor to form a pocket that is suitable for Dishevelled binding.

Besides the concerted movement of the site that is crucial for the binding of Dishevelled, two unusual and closely coupled, dynamic kinks (kink 1 and kink 2) can be seen in helix VII (Fig. 4c, d and Extended Data Fig. 9a). These two kinks are involved in conserved polar networks. At kink 1, the hydrogen-bonding network flips between two states: from S317^{3.40f}–Y444^{6.40f} and Y262^{2.51f}–S484^{7.45f} to Y444^{6.40f}–Y262^{2.51f} (Fig. 4c and Extended Data Fig. 9b). Mutation of Y444^{6.40f} in the centre of kink 1 reduced the activation of FZD4 (Fig. 4e and Extended Data Fig. 9c). Compared with kink 1, S233^{1.46f} in kink 2 is involved in a tight hydrogen-bonding network in which it keeps the lower portion of helix VII in a bent conformation (Fig. 4d and Extended Data Fig. 9b). The loss of the polar side chain of S233^{1.46f} would act to favour the linear conformation of kink 2 over the bent conformation. It is notable that S233A greatly increases Norrin-induced signalling with no effect on WNT3A signalling (Fig. 4e and Extended Data Fig. 9c). One possible explanation is that the dynamics seen at kink 2, and by extension at H8, are specific to Norrin signalling. These uniquely coupled kinks in the structure of Δ CRD-FZD4 undergo considerable movement in helix VII

and H8 during simulations (Extended Data Fig. 9a), whereas helix VI remains relatively stable. The ‘bent-to-linear’ transformation in helix VII and H8 may relate to receptor activation, which is in contrast to the outward movement of helix VI that is seen in the receptors of other family members. In addition, FZD activation requires co-receptors such as LRP5 and LRP6, and may include the rearrangement of homo- or heterodimers. Previous studies have reported that an artificial ligand that induces the formation of a heterodimer between FZD and the co-receptor LRP6 is sufficient for activating the pathway¹³. The observation that dimerization of FZD and LRP6 (by several engineered ligands) is sufficient for signalling stands against conventional ‘helix VI outward movement’ of FZDs in the activation of the Wnt pathway and is consistent with the structural insight presented here.

The structure of a Frizzled receptor transmembrane domain and the activation mechanism we present here pave the way to revealing the function of FZDs as essential gatekeepers in the Wnt signalling pathway. As emerging cancer targets, the role of FZDs remains elusive. The more structural insight that is available for each FZD, the greater the likelihood of understanding their signalling pathways and developing successful drugs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0447-x>.

Received: 12 October 2017; Accepted: 10 July 2018;
Published online 22 August 2018.

- Nusse, R. & Clevers, H. Wnt/ β -catenin signaling, disease, and emerging therapeutic modalities. *Cell* **169**, 985–999 (2017).
- Clevers, H. & Nusse, R. Wnt/ β -catenin signaling and disease. *Cell* **149**, 1192–1205 (2012).
- Dijksterhuis, J. P., Petersen, J. & Schulte, G. WNT/Frizzled signalling: receptor-ligand selectivity with focus on FZD-G protein signalling and its physiological relevance: IUPHAR Review 3. *Br. J. Pharmacol.* **171**, 1195–1209 (2014).
- Schulte, G. International union of basic and clinical pharmacology. LXXX. The class Frizzled receptors. *Pharmacol. Rev.* **62**, 632–667 (2010).
- Wang, Y., Chang, H., Rattner, A. & Nathans, J. Frizzled receptors in development and disease. *Curr. Top. Dev. Biol.* **117**, 113–139 (2016).
- Kahn, M. Can we safely target the WNT pathway? *Nat. Rev. Drug Discov.* **13**, 513–532 (2014).
- Tao, L. et al. Frizzled proteins are colonic epithelial receptors for *C. difficile* toxin B. *Nature* **538**, 350–355 (2016).
- Nile, A. H., Mukund, S., Stanger, K., Wang, W. & Hannoush, R. N. Unsaturated fatty acyl recognition by Frizzled receptors mediates dimerization upon Wnt ligand binding. *Proc. Natl Acad. Sci. USA* **114**, 4147–4152 (2017).
- DeBruine, Z. J. et al. Wnt5a promotes Frizzled-4 signalosome assembly by stabilizing cysteine-rich domain dimerization. *Genes Dev.* **31**, 916–926 (2017).
- Janda, C. Y., Waghay, D., Levin, A. M., Thomas, C. & Garcia, K. C. Structural basis of Wnt recognition by Frizzled. *Science* **337**, 59–64 (2012).
- Chang, T. H. et al. Structure and functional properties of Norrin mimic Wnt for signalling with Frizzled4, Lrp5/6, and proteoglycan. *eLife* **4**, (2015).
- Shen, G. et al. Structural basis of the Norrin–Frizzled 4 interaction. *Cell Res.* **25**, 1078–1081 (2015).
- Janda, C. Y. et al. Surrogate Wnt agonists that phenocopy canonical Wnt and β -catenin signalling. *Nature* **545**, 234–237 (2017).
- Gurney, A. et al. Wnt pathway inhibition via the targeting of Frizzled receptors results in decreased growth and tumorigenicity of human tumors. *Proc. Natl Acad. Sci. USA* **109**, 11717–11722 (2012).
- Umbhauer, M. et al. The C-terminal cytoplasmic Lys–Thr–X–X–Trp motif in Frizzled receptors mediates Wnt/ β -catenin signalling. *EMBO J.* **19**, 4944–4954 (2000).
- Byrne, E. F. X. et al. Structural basis of Smoothed regulation by its extracellular domains. *Nature* **535**, 517–522 (2016).
- Wang, C. et al. Structure of the human smoothed receptor bound to an antitumour agent. *Nature* **497**, 338–343 (2013).
- Zhang, X. et al. Crystal structure of a multi-domain human smoothed receptor in complex with a super stabilizing ligand. *Nat. Commun.* **8**, 15383 (2017).
- Huang, P. et al. Structural basis of Smoothed activation in Hedgehog signaling. *Cell* **174**, 312–324.e16 (2018).
- Wang, Y. et al. Norrin/Frizzled4 signaling in retinal vascular development and blood brain barrier plasticity. *Cell* **151**, 1332–1344 (2012).
- Zhang, C. et al. Norrin-induced Frizzled4 endocytosis and endo-lysosomal trafficking control retinal angiogenesis and barrier function. *Nat. Commun.* **8**, 16050 (2017).
- Nikopoulos, K. et al. Overview of the mutation spectrum in familial exudative vitreoretinopathy and Norrie disease with identification of 21 novel variants in *FZD4*, *LRP5*, and *NDP*. *Hum. Mutat.* **31**, 656–666 (2010).
- Strakova, K. et al. The tyrosine Y250²³⁹ in Frizzled 4 defines a conserved motif important for structural integrity of the receptor and recruitment of Dishevelled. *Cell. Signal.* **38**, 85–96 (2017).
- Zhang, H. et al. Structure of the full-length glucagon class B G-protein-coupled receptor. *Nature* **546**, 259–264 (2017).
- Jazayeri, A. et al. Crystal structure of the GLP-1 receptor bound to a peptide agonist. *Nature* **546**, 254–258 (2017).
- Wu, H. et al. Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science* **344**, 58–64 (2014).
- Zheng, Y. et al. Structure of CC chemokine receptor 2 with orthosteric and allosteric antagonists. *Nature* **540**, 458–461 (2016).
- Rasmussen, S. G. et al. Crystal structure of the β_2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011).

Acknowledgements This work was supported by the National Natural Science Foundation (NSF) of China grant 31670736 (F.X.), the National Key Research and Development Program of China grant 2018YFA0507004 (F.X.) and 2016YCF0905902 (S.Z.), the NSF of Shanghai grant 16ZR1448500 (S.Z.), the Russian Foundation for Basic Research (RFBR 18-34-00990) (P.P.) and Shanghai Municipal Government, ShanghaiTech University. The diffraction data were collected at BL41XU@Spring-8 with JASRI proposals 2016B2702. We thank J. Liu, X. Gu, N. Chen and L. Xue of the BV facility at the iHuman Institute, ShanghaiTech University for protein expression support; M. Hanson from the GPCR Consortium and K. Diederichs from the University of Konstanz for X-ray data processing; the mass spectrometry facility at the National Protein Science Center (Shanghai, China) for technical assistance; and A. Pautsch from Boehringer Ingelheim and W. Zhong from Amgen for discussions.

Reviewer information *Nature* thanks M. Filizola, X. He, A. K. Shukla and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.Y. performed cloning, protein purification, crystallization, data collection, structural analysis and figure preparation; Y.W. carried out structure analysis, molecular dynamics simulations and figure preparation; T.-H.X. performed mutagenesis, cellular localization, cell surface biotinylation, and TOPflash reporter assays and corresponding figure preparation; P.W.d.W. carried out molecular dynamics simulations of Δ CRD–FZD4, structure analysis and corresponding figure preparation; Y.H., Z.J.D., Y.L., K.S.-P. and K.G.H. performed mutagenesis and TOPflash reporter assays; M.P. carried out molecular replacement and structure refinement; B.Z. carried out affinity mass spectrometry; S.A.Z. performed full-length modelling; P.P. was responsible for stabilizing mutation design; G.W.H. was responsible for structure refinement, quality control and deposition; Y.C. and S.D. performed cloning and protein purification; Y.G. carried out computational analysis; V.K. supervised the stabilization of mutation design and full-length modelling; W.S. supervised the affinity mass spectrometry analysis and table preparation; L.J.M., K.M. and H.E.X. designed cell-based experiments, data analysis and interpretations; R.C.S. supervised the structure analysis; S.Z. supervised the structure analysis, simulation and figure preparation; F.X. designed and supervised experiments and performed data analysis; and S.Y. and F.X. wrote the manuscript with discussions and improvements from Y.W., P.W.d.W., H.E.X., K.M., S.Z. and R.C.S.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0447-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0447-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to F.X.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Purification of Δ CRD-FZD4. The construct of Δ CRD-FZD4 (residues 178–517) was designed with four mutations (M309L, C450I, C507F, S508Y) based on computational predictions of stabilizing mutations made with CompoMug tool²⁹ and with eight residues (420–427) at ICL3 replaced by rubredoxin. This construct was expressed in *Spodoptera frugiperda* (Sf9) cells with N-terminal haemagglutinin signal peptide, Flag tag and 10 \times His tag as previously described³⁰. The cell membrane was washed with a low-salt buffer containing 10 mM HEPES (pH 7.5), 20 mM KCl, 10 mM MgCl₂ and protease inhibitor cocktail (Roche). Then the membrane was washed with a high-salt buffer containing 10 mM HEPES (pH 7.5), 1 M NaCl, 20 mM KCl, 10 mM MgCl₂ and protease inhibitor cocktail. To avoid aggregation by free cysteine, the membrane was incubated with 2 mg ml⁻¹ iodoacetamide (Sigma) at 4°C for 30 min. After incubation, the membrane solution was incubated with solubilization buffer containing 10 mM HEPES (pH 7.5), 20 mM KCl, 10 mM MgCl₂ and 2% (w/v) lauryl maltose neopentyl glycol (LMNG; Anatrace) at a ratio of 1:1 at 4°C for 1 h. The sample was centrifuged at 160,000g for 40 min to remove debris, and the supernatant was incubated with TALON IMAC resin (Clontech) at 4°C overnight. Then, the resin was washed with 50 column volumes of 50 mM HEPES (pH 7.5), 500 mM NaCl, 10% (v/v) glycerol, 1% (w/v) LMNG, 10 mM MgCl₂, 8 mM ATP and 10 mM imidazole, followed by 6 column volumes of 50 mM HEPES (pH 7.5), 500 mM NaCl, 10% (v/v) glycerol, 0.05% (w/v) LMNG and 40 mM imidazole. Fluorescent dye Cy3 NHS ester (GE Healthcare) was then added to the resin to a final concentration of 10 μ M (to monitor the protein crystals). After incubation at 4°C for 2 h, free fluorescent dye was removed by washing with 200 column volumes of 25 mM HEPES (pH 7.5), 500 mM NaCl, 10% (v/v) glycerol and 0.01% (w/v) LMNG. The receptor was then eluted with 25 mM HEPES (pH 7.5), 500 mM NaCl, 10% (v/v) glycerol, 0.01% (w/v) LMNG and 200 mM imidazole. Imidazole was removed using a PD MiniTrap G-25 column (GE Healthcare).

Crystallization of Δ CRD-FZD4. The receptor was concentrated to about 40 mg ml⁻¹ with a 50-KDa cutoff concentrator (Millipore), and reconstituted into the lipidic cubic phase by mixing with monoolein at a protein/lipid ratio of 1:1.2. The crystallization trials were set up by crystallization robot NT8 (Formulatrix). The crystals emerged under the conditions of 100 mM sodium cacodylate trihydrate (pH 6.0), 80 mM MgSO₄, 30% PEG400 and 1.5–2.5% v/v (\pm)-2-methyl-2,4-pentanediol at 3 days in the sponge phase.

Data collection and structure determination. The X-ray diffraction data were collected at Spring-8 beam line 41XU, Hyogo, Japan, using a PILATUS detector (X-ray wavelength 1.0000 Å). A rastering system was used to find the best diffracting region of single crystals³¹. The crystals were exposed for 0.2 s and 0.2° oscillation per frame. XDS³² was used for integrating and scaling data from the 36 crystals. A molecular replacement method with Phaser³³ was applied to obtain the initial phase information using the structures of the receptor portion of SMO (PDB ID: 4JKV) and of rubredoxin (PDB ID: 1IRN) as search models. The refinement was performed with Phenix³⁴ and Buster³⁵ followed by manual examination and rebuilding of the refined coordinates in Coot³⁶ using both $|2F_o| - |F_c|$ and $|F_o| - |F_c|$ maps. Ramachandran plot analysis of the final structures with MolProbity showed that 100% of the residues are in either favoured (95.2%) or allowed (4.8%) regions, with no outlier. The final data collection and refinement statistics are shown in Extended Data Table 1a.

Modelling and molecular dynamics simulations of full-length FZD4. Processing of the protein structure was performed with the Protein Preparation Wizard tool³⁷ in Schrödinger Suite 2015-4. ICL3 (cut in the construct) was built using the Prime tool³⁸ in Schrödinger. Calculations of pocket volume and cross-sectional area were performed using the program Cavity³⁹. The positions of receptors in the membrane were obtained from the OPM database⁴⁰. Amino acid conservation scores were obtained from the ConSurf server⁴¹. Generation of conservation scores, worm representation and calculations of hydrophilic/hydrophobic surface area ratio (based on non-carbon and carbon atoms) were performed with UCSF Chimera⁴². A full-length model of FZD4 was built using a multi-template homology modelling tool implemented in ICM-Pro (Molsoft). The model used the current structure of FZD4 TMD and CRD (PDB: 5CM4) structurally aligned to the full-length structure of SMO (PDB: 5L7D). The FZD4 polypeptide chain (residues 44–513) was threaded through the TMD and CRD structures and thoroughly optimized by conformational sampling.

Molecular dynamics simulations were performed with GROMACS 5.1.2⁴³ using force field CHARMM36⁴⁴. Crystal structures of FZD4 were embedded into a pre-equilibrated POPC (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylcholine) lipid bilayer with water using the membed tool in the program GROMACS. Sodium ions were added to a concentration of 0.15 M in water, and chloride ions were added to neutralize the system. Molecular dynamics simulations were run independently three times. First, atom velocity was generated at a temperature of 310 K. Then the system was relaxed in a canonical (NVT) ensemble for 300 ps and balanced in position-restrained molecular dynamics (isothermal–isobaric

(NPT) ensemble with pressure of 1 atm, using semi-isotropic coupling) for 15 ns (total energy was stable). Finally, productive molecular dynamics with no position restraints was run for 50 ns. The coordinates of atoms were recorded every 1 ps. The input files for molecular dynamics simulations of the full-length FZD4 models were generated using the CHARMM-GUI⁴⁵ server, and SMO-derived lipid coordinates obtained from the OPM⁴⁶ server were used to embed the full-length model in the POPC layer. In three independent runs, the system was equilibrated under the NVT ensemble for 500 ps (1-fs time-step) and the NPT for 15 ns (2-fs time-step) with stepwise reduction in positional restraints on protein atoms, before the final production run of 1,500 ns.

Molecular dynamics simulations of Δ CRD-FZD4. All-atom atmospheric simulations of Δ CRD-FZD4 were performed using GROMACS5.0.6⁴³ in the NPT ensemble with periodic boundary conditions and the CHARMM36m force field⁴⁷. The receptor was prepared for simulation by removing all heteroatoms and the rubredoxin fusion partner. Second, thermostabilizing mutants were reverted back to wild type with Modeller 9.18⁴⁸ and aligned for membrane insertion using the Orientations of Proteins database PPM server⁴⁶. Titratable residues were left in their dominant state at pH 7.0 and all histidines were represented with a hydrogen on the epsilon nitrogen. The resulting Δ CRD-FZD4 was capped with neutral acetyl and methylamine groups and embedded into a pre-equilibrated POPC lipid bilayer solvated in a box of TIP3P waters allowing for at least 14 Å of padding on all sides with 150 mM NaCl, and neutralised by removing appropriate ions or counter ions using the Desmond system builder within Maestro (Schrödinger Release 2018-1: Maestro, Schrödinger). Final system dimensions were 96 \times 96 \times 114 Å and the system comprised 185 lipids, 16,722 water molecules, 46 chloride ions and 48 sodium ions.

Before production simulations, 50,000 steps of energy minimization were performed followed by equilibration in the NVT and NPT ensembles for 10 and 50 ns, respectively, with positional restraints (1,000 kJ mol⁻¹ nm⁻²) placed on heavy atoms. A second round of NTP equilibration for 50 ns was run with positional restraints (1,000 kJ mol⁻¹ nm⁻²) on backbone atoms to allow for sidechain relaxation. System temperature was maintained at 310 K using the v-rescale method with a coupling time of 0.1 ps and pressure was maintained at 1 bar using the Berendsen barostat with a coupling time of 1.0 ps and compressibility of 4.5 \times 10⁻⁵ bar⁻¹ with semi-isotropic coupling. Simulations were performed with a 2-fs timestep and all bond lengths were constrained using LINCS. Electrostatic interactions were computed using the particle mesh Ewald (PME) method with non-bonded interactions cut at 10.0 Å.

Three independent 3- μ s production simulations of Δ CRD-FZD4 were performed using the Parrinello–Rahman barostat with a coupling time of 5.0 ps for a combined total of 9 μ s. During production, trajectory snapshots were saved every 10 ps. Simulation analysis was performed using MDTraj 1.7.2⁴⁹ and VMD 1.9.2⁵⁰. Plots were generated using the R statistical package (<http://www.R-project.org>). System parameters and trajectories are available upon request.

Cell-based luciferase assay. Mutations were introduced by QuickChange site-directed mutagenesis (Stratagene). All constructs and mutations were sequence-verified. HEK293 cells were maintained in DMEM (Gibco) with 5% fetal bovine serum. Cell-based luciferase assays were performed as previously described⁵¹ with small modifications. In brief, cells were split at 20,000 per well in a 24-well plate 24 h before transfection. Cells were transfected with 10 ng FZD4-YFP expression plasmid, 10 ng LRP6 expression plasmid, 100 ng SuperTOPflash TCF-luciferase reporter, 10 ng ligand expression vector (WNT3A or Norrin or vehicle) and 1 ng pRGtkRenilla control plasmid using X-tremeGene 9 (Roche) per well. Cells were collected and lysed 48 h after transfection. Luminescence activities were measured using the Dual Luciferase Kit (Promega) according to the manufacturer's instructions. Renilla luciferase serves as transfection control. All activities were normalized (relative activities) to the activity of basal wild-type FZD, which was set as 1.0.

Fluorescence microscopy and protein expression. Wild-type and mutant FZD4-YFP fluorescence were visualized in non-stimulated cells 48 h after transfection using a Nikon Eclipse TE300 mercury lamp microscope. YFP fluorescence was excited at 488 nm, and emission was detected at 496–518 nm with 5-s exposure. Each corresponding YFP view was also taken in bright field with 200-ms exposure. We used a plasmid expressing C-terminally sfGFP-tagged arrestin protein, which localizes largely to the cytoplasm, as negative control for cell membrane localization⁵².

After fluorescence images were taken, the cells were lysed in CellLytic M (Sigma-Aldrich) with 1 \times protease inhibitor mixture (Roche Diagnostics) for SDS–PAGE analysis. FZD4-YFP bands were imaged using a ChemiDoc MP Imaging System (Bio-Rad) with Alexa 488 filter setting and 600-s exposure time.

Cell surface biotinylation assay. Cell surface biotinylation assays were carried out as previously described⁵². In brief, HEK293 cells were seeded at a density of 0.8 \times 10⁶ per ml in 6-well plates and transfected the following day with 0.5 μ g of FZD4-YFP expression plasmid and 0.5 μ g LRP6 expression vector with

Lipofectamine 2000 (Invitrogen) transfection reagent. After 48 h, plates were placed on ice, medium was carefully aspirated and cells were washed with cold PBS (20 mM potassium phosphate, pH 7.4, 150 mM NaCl). The cells were then incubated with 1 ml per well of PBS with 0.25 mg ml⁻¹ EZlink-Sulfo-NHS-SS-biotin (Pierce) for 40 min at 4°C. Biotinylation was stopped by the addition of 100 mM glycine in PBS (1 ml). Cells were washed once with cold TBS and lysed in CellLytic M (Sigma-Aldrich) with 1× protease inhibitor mixture (Roche Diagnostics). Non-solubilized material was removed by centrifugation (10 min at 20,000g and 4°C). 20 µl of supernatant plus 20 µl of 2× SDS loading buffer served as input. The remaining supernatant was incubated with prewashed Streptavidin MagBeads (GenScript) for about 30 min at 4°C, followed by three washes with lysate buffer. Biotinylated proteins were eluted by reduction of the NHS-SS-biotin bond with SDS loading buffer containing 100 mM dithiothreitol for 20 min at 65°C. Eluates were subjected to SDS-PAGE for fluorescent imaging and western blot analysis using anti-actin or anti-α-tubulin antibodies as loading control. For input samples, the fluorescent images were exposed for 30 s and for elute samples for 600 s.

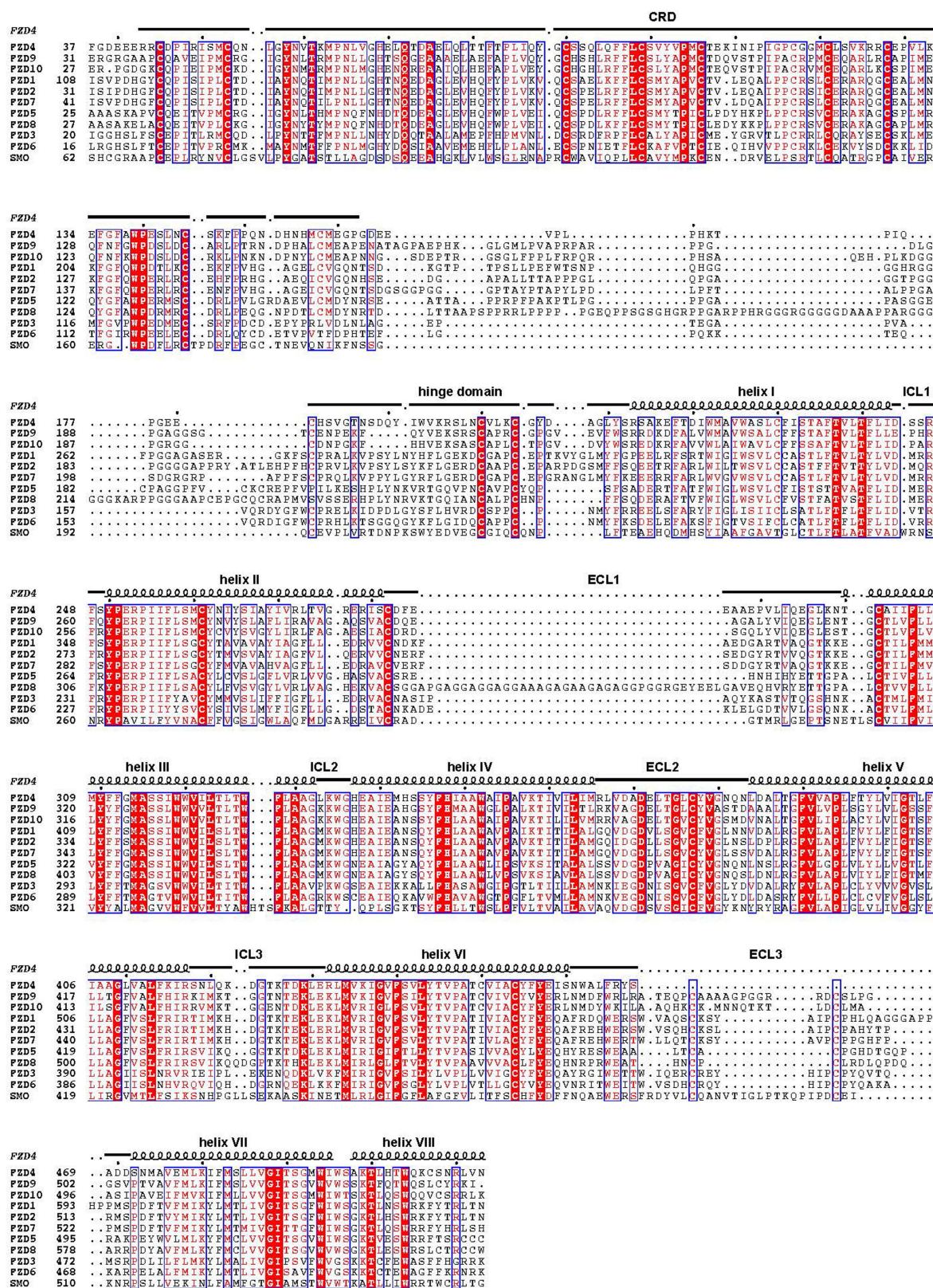
Detection of receptor and ligand interaction with affinity mass spectrometry analysis. The purified FZD4 or SMO proteins were incubated with LY2940680 at a final concentration of 250 nM (protein) and 50 nM (compound) at 4°C for 60 min. Then the ligand-bound protein complexes were separated from free ligands by ultracentrifugation as previously described⁵³. Purified A2a receptor that underwent the same process served as a negative control. Compounds released from protein complexes were analysed by an Agilent 6530 TOF system equipped with Agilent 1260 HPLC with a reported method⁵³. Experimental triplicates were prepared for each pair of the receptor of interest and the negative control. Responses of specific compounds were extracted using MassHunter software (Agilent) based on the accurate mass measurement (<10 p.p.m. error) and matching retention time of the compound standard (<0.1 min shift). S/C ratios refer to the ratio of the mass spectrometry response of a specific compound detected in the protein incubation sample compared with the control. Average S/C > 2 (r.s.d. <30%) or presence of the compound in the protein incubation sample only indicates positive binding of the compound to the protein target⁵³.

Statistics and reproducibility. For TOPflash reporter assays, each data point was determined from three independent transfectants ($n = 3$), represented as mean ± s.e.m. Sample sizes were selected by power analysis for a 95% confidence interval to detect a 1.5-fold change relative to wild type, for a standard deviation of wild-type values of 0.1 and mutant values of 0.2 (statistical power = 98.7%). Cell surface expression data were obtained using two independent methods—fluorescence microscopy and cell surface biotinylation—with similar results, and a subset of the fluorescence microscopy data was further validated independently by another laboratory member. The majority of transfections and TOPflash assays were independently repeated at least once in triplicate with similar results. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

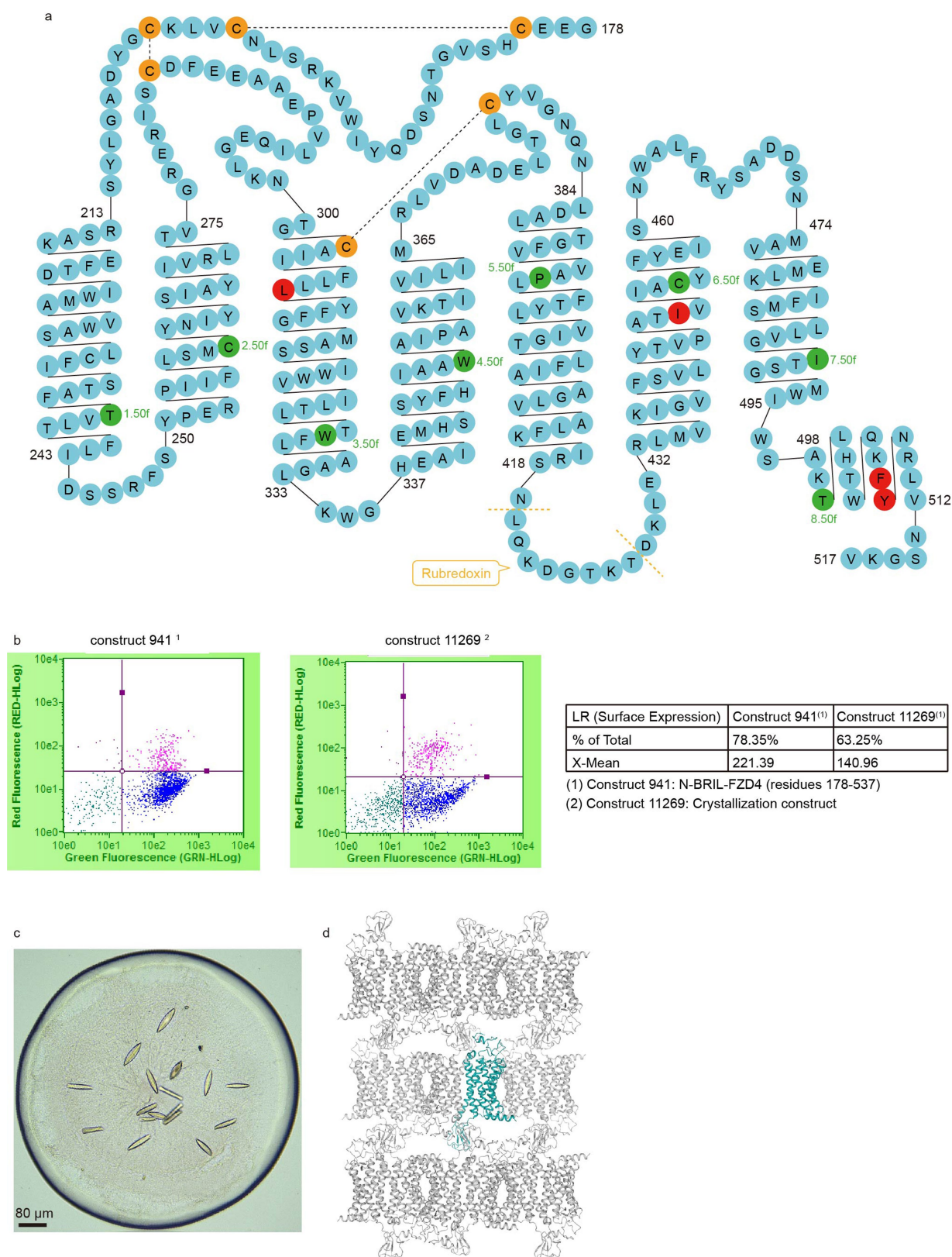
Data availability. Coordinates and structure factors for ΔCRD-FZD4 have been deposited in the Protein Data Bank (PDB) with the accession number 6BD4. All other data relating to this study are available from the corresponding author on reasonable request.

29. Popov, P. et al. Computational design of thermostabilizing point mutations for G protein-coupled receptors. *eLife* **7**, e34729 (2018).
30. Ma, Y. et al. Structural basis for apelin control of the human apelin receptor. *Structure* **25**, 858–866.e4 (2017).
31. Cherezov, V. et al. Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-µm size X-ray synchrotron beam. *J. R. Soc. Interface* **6**, S587–S597 (2009).
32. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
33. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
34. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
35. Smart, O. S. et al. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. D* **68**, 368–380 (2012).
36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
37. Pau, M. S., Gao, S., Malbon, C. C., Wang, H. Y. & Bertalovitz, A. C. The intracellular loop 2 F328S Frizzled-4 mutation implicated in familial exudative vitreoretinopathy impairs Dishevelled recruitment. *J. Mol. Signal.* **10**, 5 (2015).
38. Kramer, G. D., Say, E. A. & Shields, C. L. Simultaneous novel mutations of LRP5 and TSPAN12 in a case of familial exudative vitreoretinopathy. *J. Pediatr. Ophthalmol. Strabismus* **53**, e1–e5 (2016).
39. Yuan, Y., Pei, J. & Lai, L. LigBuilder 2: a practical *de novo* drug design approach. *J. Chem. Inf. Model.* **51**, 1083–1091 (2011).
40. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of Proteins in Membranes database. *Bioinformatics* **22**, 623–625 (2006).
41. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
42. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
43. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
44. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
45. Lee, J. et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).
46. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
47. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
48. Eswar, N. et al. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **15**, 5.6.1–5.6.30 (2006).
49. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
50. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **14**, 33–38 (1996).
51. Xu, T. H. et al. Alzheimer's disease-associated mutations increase amyloid precursor protein resistance to γ-secretase cleavage and the Aβ42/Aβ40 ratio. *Cell Discov.* **2**, 16026 (2016).
52. Yan, Y. et al. Dimerization of the transmembrane domain of amyloid precursor protein is determined by residues around the γ-secretase cleavage sites. *J. Biol. Chem.* **292**, 15826–15837 (2017).
53. Chen, X. et al. A ligand-observed mass spectrometry approach integrated into the fragment based lead discovery pipeline. *Sci. Rep.* **5**, 8361 (2015).



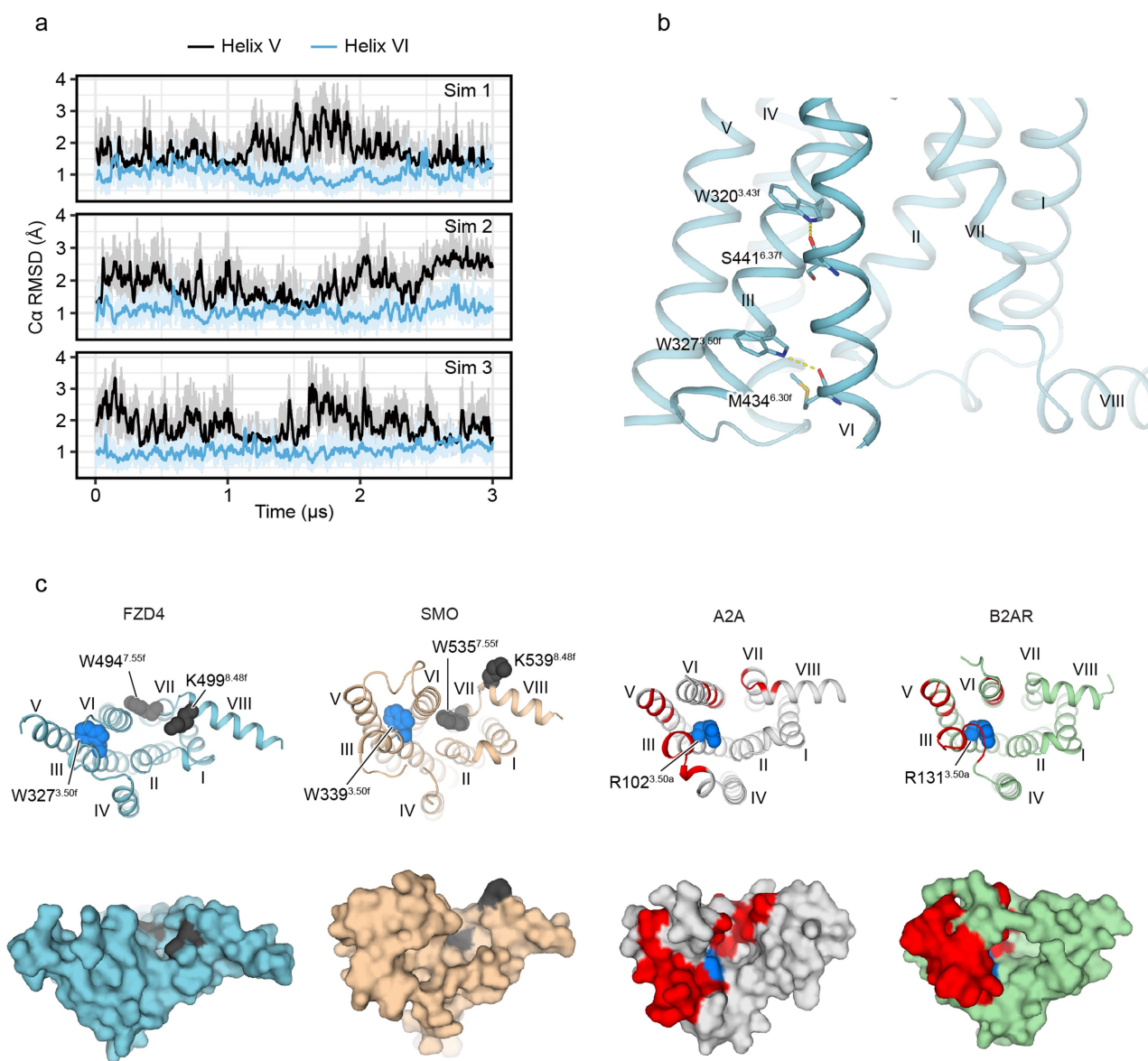
Extended Data Fig. 1 | Sequence alignment between FZD4 and nine other FZDs as well as SMO. Colours represent the similarity of residues: red background, identical; red text, strongly similar. The alignment was

generated using MAFFT (<https://www.ebi.ac.uk/Tools/msa/mafft/>) and the graphic was prepared on the ESPrnt 3.0 server (<http://esprnt.ibcp.fr/ESPrnt/cgi-bin/ESPrnt.cgi>).



Extended Data Fig. 2 | Crystallization of Δ CRD-FZD4 and structure determination. **a**, Schematic of the Δ CRD-FZD4 construct. To obtain crystals that would diffract well, we truncated the N-terminal CRD region (residues 1–177) and C-terminal flexible region (residues 517–537) and introduced four single mutations (M309L, C450I, C507F and S508Y, coloured in red) that are designed based on sequence conservation analysis across ten human FZDs. Residues that are involved in the X.50f numbering

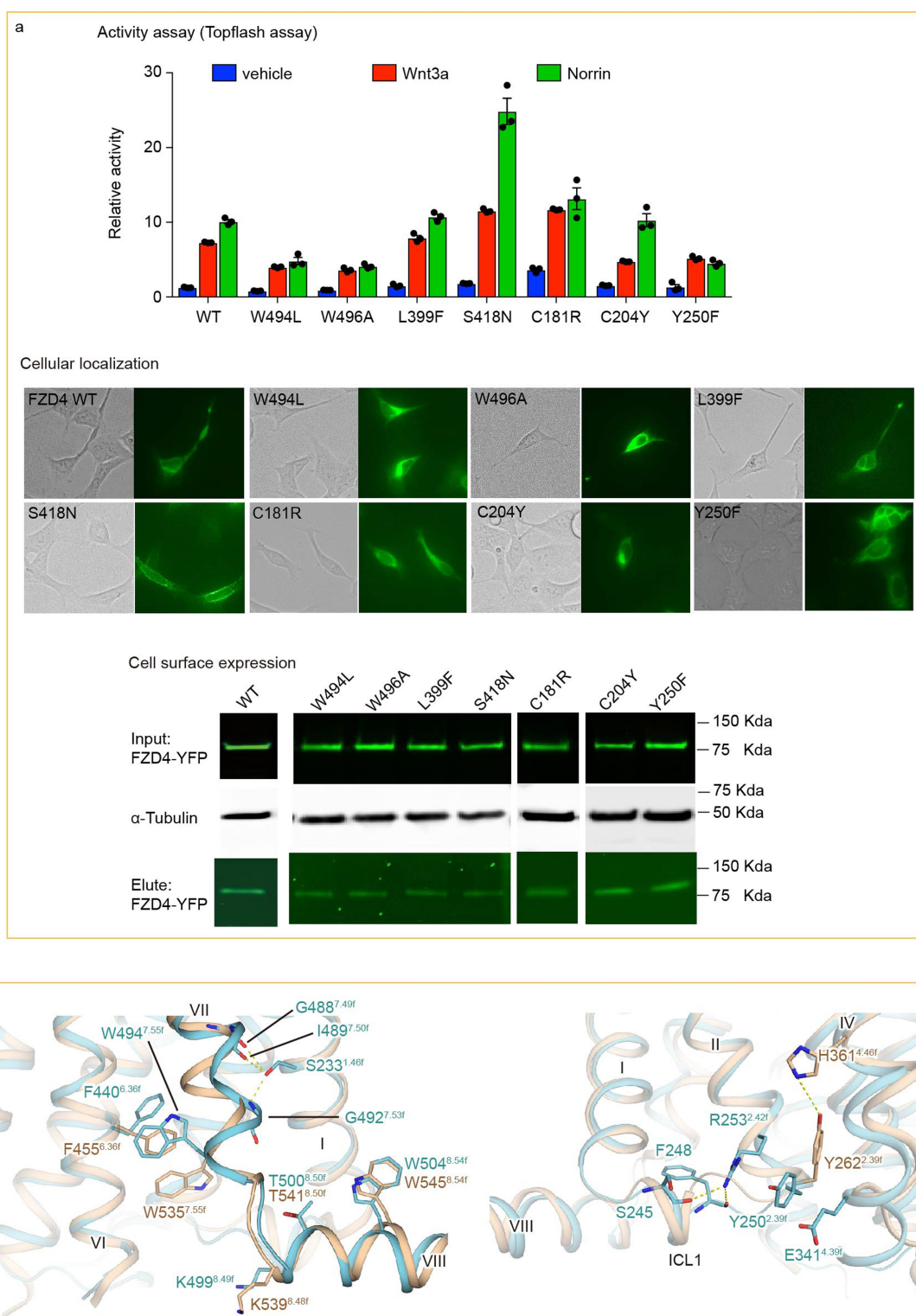
system are coloured in green. The cysteines that form endogenous disulfide bonds are indicated in orange. **b**, Fluorescence-activated cell sorting staining data, to monitor the surface expression of the construct used in this study. The experiment was repeated twice with similar results. **c**, Crystals of Δ CRD-FZD4 in the apo state. **d**, Crystal packing of Δ CRD-FZD4.



Extended Data Fig. 3 | Dynamics of helix V and VI in FZD4 and a comparison of the intracellular side of FZD4 with other GPCRs.

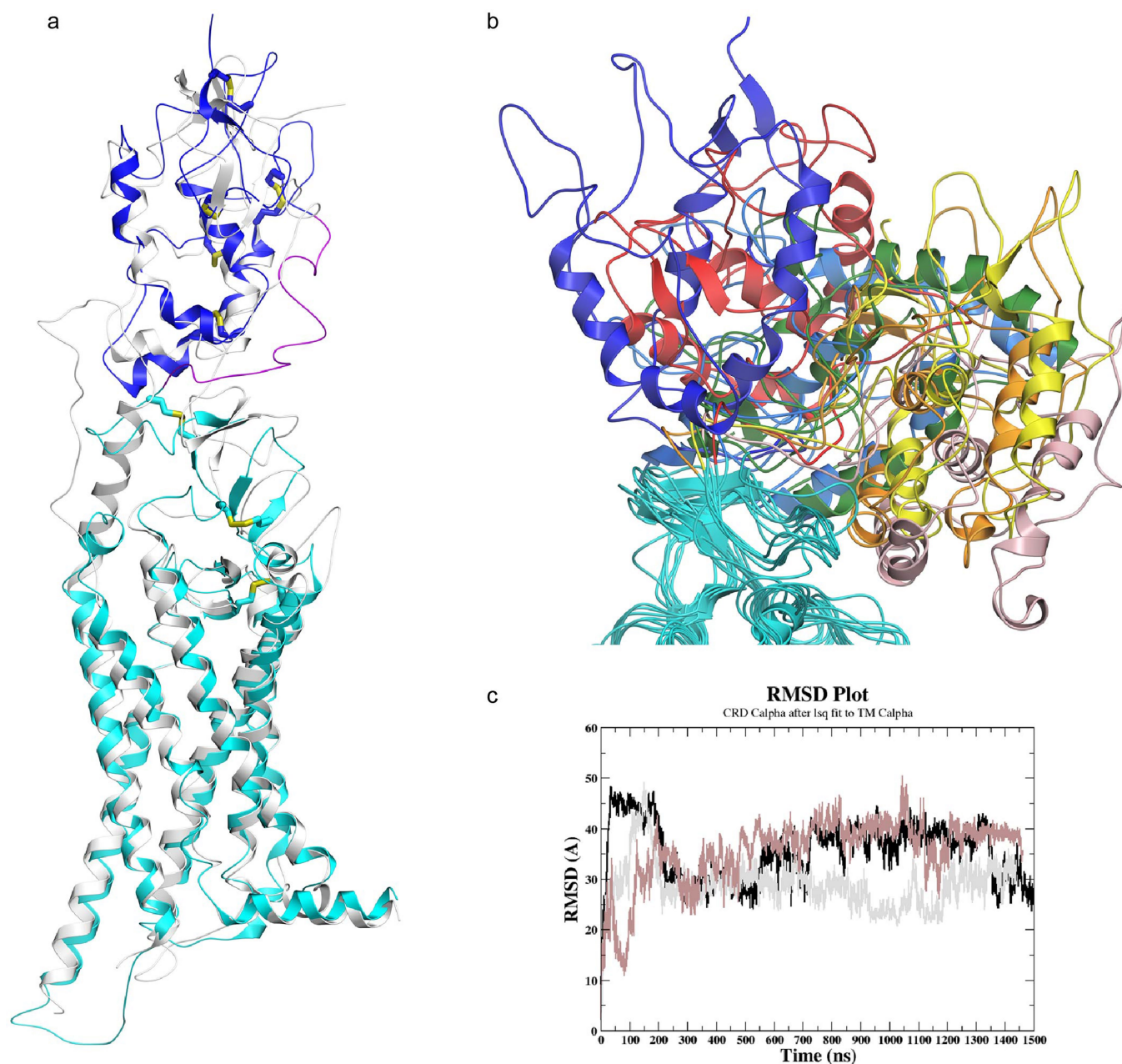
a, C α r.m.s.d. for residues G409 to S418 (helix V) and R432 to S441 (helix VI) plotted as a rolling average over a 10-ns window. **b**, Unique hydrogen bonds between helix III and helix VI (W320^{3.43f}–S441^{6.37f} and W327^{3.50f}–M434^{6.30f}) maintain helix VI in an inward conformation. **c**, The intracellular cavity of FZD4 is close to W^{7.55f} and the KTXXXW motif (coloured in dark grey), a region that is key for downstream signalling.

Compared with class-A GPCRs (adenosine A_{2A} and β 2 adrenergic receptors), FZD4 leaves no cavity among helices III, VI and VII, whereas SMO has a side-pocket at this position. R^{3.50} (labelled with a blue sphere) is a key residue in the activation of class-A GPCRs and is exposed to the intracellular surface. The residue in the same position in class-F, W^{3.50f}, points in a different direction and is not exposed. The G-protein-binding sites of class-A GPCRs are labelled in red.



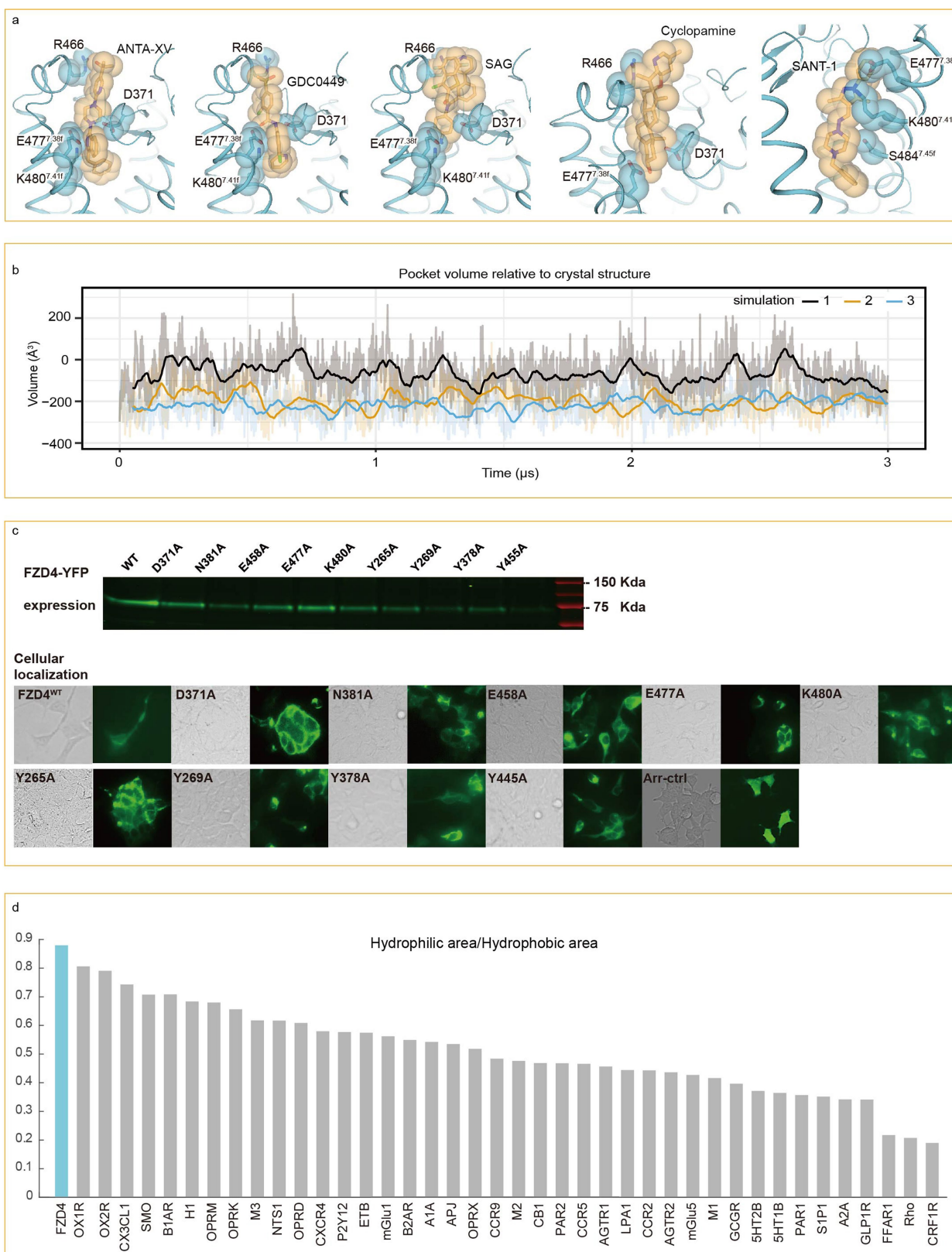
Extended Data Fig. 4 | Activity analysis of disease-related and signalling-related mutations of Δ CRD-FZD4, and the conformational change of family-conserved amino acids of Δ CRD-FZD4. **a, TOPflash analysis, cellular localization and cell surface expression analysis of disease-related and signalling-related mutations. Each data point represents mean \pm s.e.m., repeated in triplicate. The experiment was repeated using two independent methods with similar results. All the**

mutations affected FZD4 downstream signalling in some way—some decreased the WNT3A–Norrin signal substantially, some showed gain-of-function activity—which suggests complex mechanisms underlying these mutation-caused diseases. **b**, Conformational rearrangement on family-conserved residues W494^{7.55f} and Y250^{2.39f} was observed when comparing FZD4 (cyan) with SMO (gold). It is noteworthy that Y250^{2.39f} in FZD4 points outwards from the 7TM bundle.



Extended Data Fig. 5 | The model of full-length FZD4 and molecular dynamics analysis. **a**, The model of full-length FZD4. Cyan, TMD (crystal structure); magenta, linker; blue, CRD (modelled from PDB ID: 5CM4). The human smoothened receptor crystal structure (grey; PDB ID: 5L7D) is overlaid onto the model. Disulfide bridges in the FZD4 model are shown in stick representation. **b**, Motions of the CRD domain observed during the molecular dynamics simulation. Snapshots of the positions of the CRD domain and linker during one of three independent molecular dynamics

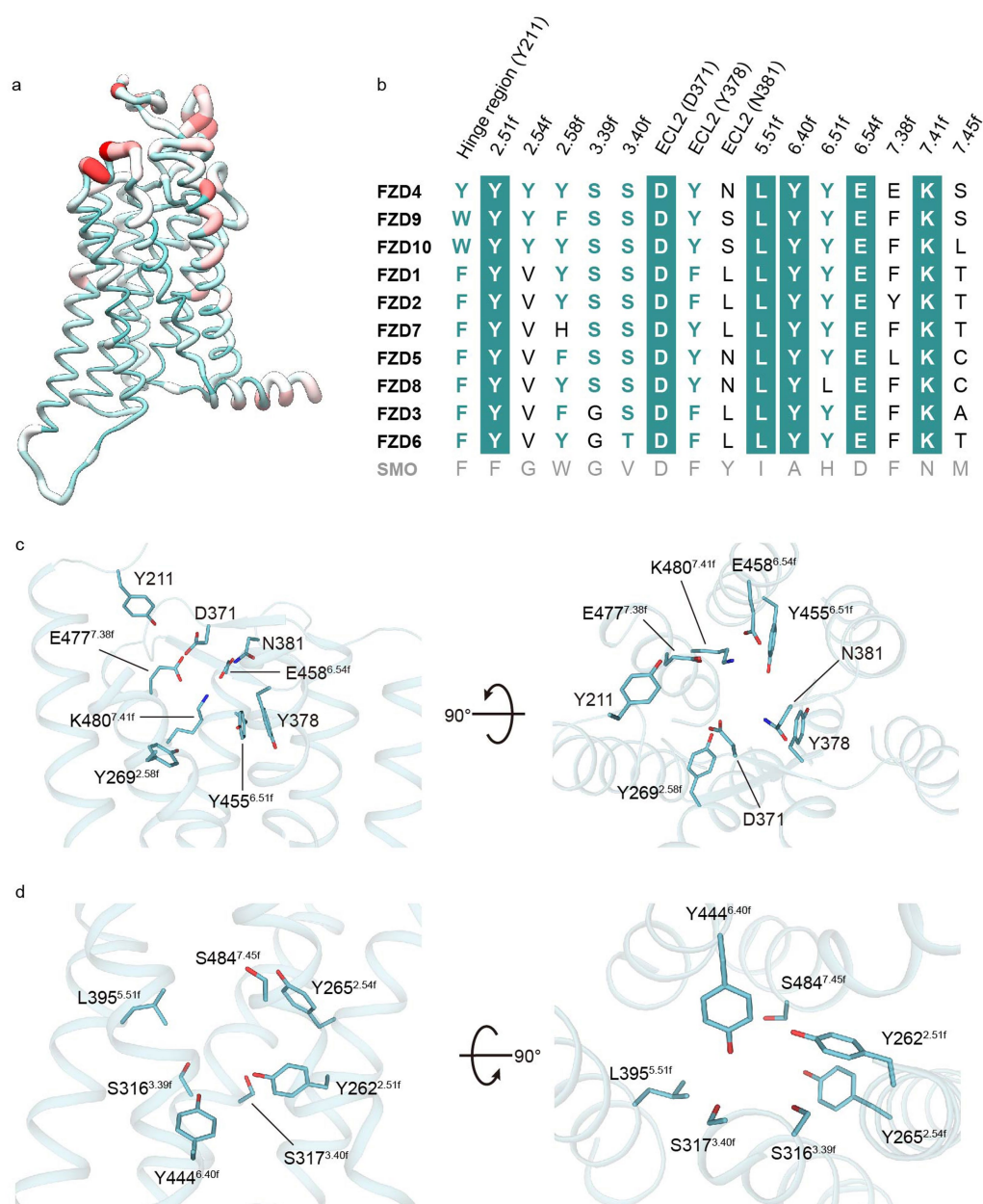
simulations are shown in cartoon representation, at 250 ns (light blue), 500 ns (green), 750 ns (pink), 1,000 ns (yellow), 1,250 ns (orange) and 1,500 ns (red). The initial full-length CRD and linker domain are shown in blue, and all TMDs are shown in cyan. **c**, Motions of the CRD domain observed during the molecular dynamics simulation. The r.m.s.d. plot of C α atoms of the CRD and linker domain with respect to the initial full-length FZD4 model during the three molecular dynamics simulations is shown.



Extended Data Fig. 6 | Analysis of the TMD pocket of FZD4.

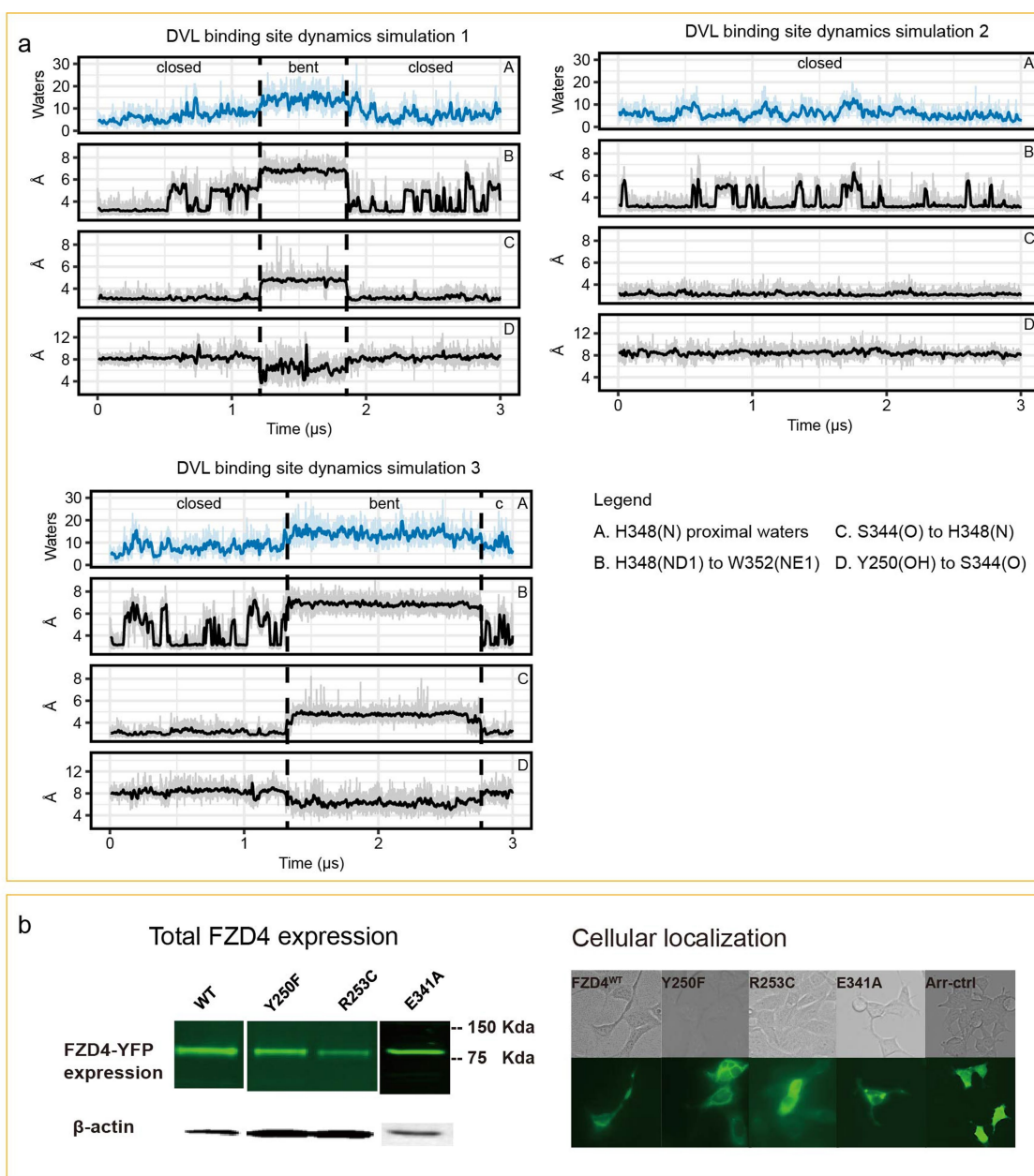
a, Superimposition of other SMO ligands (ANTA-XV, GDC0449, SAG, Cyclopamine and SANT-1) in the pocket of FZD4. All of these SMO ligands collide with the FZD4 pocket. **b**, The volume of the TMD pocket during simulation. Volumes are displayed as rolling averages over a 10-ns window. **c**, Total expression and cellular-localization analysis for pocket mutations. Total FZD-YFP expression is determined by YFP fluorescence;

cellular localization is determined by fluorescence microscopy (related to Fig. 3). A plasmid expressing C-terminally sfGFP-tagged arrestin protein (Arr-ctrl), which localizes largely to the cytoplasm, was used as negative control for cell membrane localization. The experiment was repeated using two independent methods with similar results. **d**, The FZD4 pocket has the highest hydrophilic/hydrophobic ratio of all GPCR structures that have been solved to date.



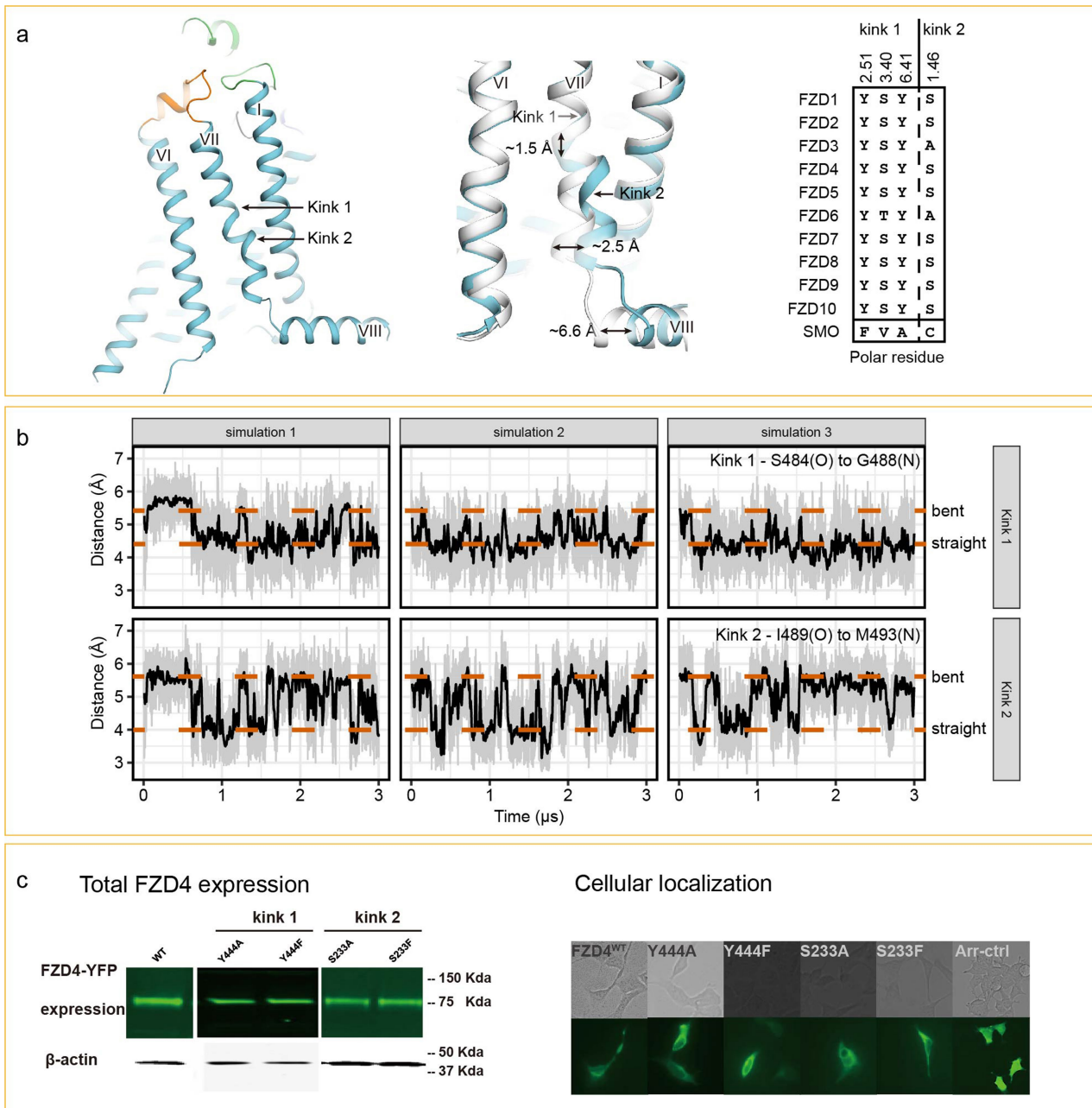
Extended Data Fig. 7 | Homology model and conservation analysis for ten human FZDs. **a**, Superposition of ten human FZD homology models. The green, thin regions represents high conservation, and the red, thick regions represent low conservation. **b**, Sequence alignment of pocket

residues, with conserved amino acids across the FZD family highlighted in dark green. **c**, Conserved amino acids in the top of the FZD4 pocket. **d**, Conserved amino acids in the bottom of the FZD4 pocket.



Extended Data Fig. 8 | Molecular dynamics and mutation analysis of the Dishevelled-binding site. a, The Dishevelled-binding site plotted as a moving average over a 10-ns window. Water molecules less than 10 Å from H348(N) were defined as proximal for this analysis. **b,** Total expression

and cellular localization analysis for mutations in the Dishevelled-binding site (related to Fig. 4). The experiment was repeated using two independent methods with similar results.



Extended Data Fig. 9 | Analysis of the two unusual kinks. a, Two kinks with conserved polar networks fluctuate between bent and straight conformations during simulation. **b**, Molecular dynamics traces of the kink 1 and kink 2 backbone distances plotted as a moving average over a

10-ns window. **c**, Total expression and cellular localization analysis of kink 1/kink 2 mutations (related to Fig. 4). The experiment was repeated using two independent methods with similar results.

Extended Data Table 1 | Crystallographic data table and affinity mass spectrometry analysis

a

	Δ CRD-FZD4
Data collection	
Space group	C222 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	61.67, 154.69, 114.40
α , β , γ (°)	90.0, 90.0, 90.0
Resolution (Å)	45.99–2.40 (2.49–2.40) ^a
<i>R</i> _{sym} or <i>R</i> _{merge}	8.54 (51.15)
<i>I</i> / σ <i>I</i>	18.19 (3.52)
Completeness (%)	100 (100)
Redundancy	20.7 (13.2)
Refinement	
Resolution (Å)	30.00–2.40
No. reflections	21,728
<i>R</i> _{work} / <i>R</i> _{free} (%)	21.00/23.30
No. atoms	
Protein	2,996
Lipids and other	205
Wilson <i>B</i> -factors (Å ²)	76.8
<i>B</i> -factors (Å ²)	
Δ CRD-FZD4	90.4
Rubredoxin	111.3
Lipids and other	110.4
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	0.76

b

	SMO	FZD4 5068 ^a	FZD4 11269 ^b
LY2940680	99 ^c	0 ^d	0

a, Data collection and refinement statistics. Values in parentheses are for highest-resolution shell. A total of 36 crystals were used. **b**, Affinity mass spectrometry analysis of the LY2940680 interaction with SMO or FZD4.

^aThe FZD4 construct with N-terminal fragment (residues 1–177) replaced by BRIL.

^bThe FZD4 construct used for crystallization in this study.

^cS/C = 99 indicates that the compound was only detected in the protein incubation sample while absent in the control, thus it specifically interacted with the corresponding receptor.

^dS/C = 0 indicates that the compound was absent in the protein incubation sample, thus it did not interact with the corresponding receptor.

CAREERS

COLLABORATION Tips from lab heads on how to support teamwork **p.673**

REPUTATION Raising the profile of female scientists on Wikipedia **go.nature.com/wiki**

LAB LIFE How to support undergraduate researchers **go.nature.com/undergraduate**



A protest in Brussels ahead of the European Union's vote to ban neonicotinoid pesticides in April 2018.

POLICY

How your science can shape policy

Decision-makers need researchers' input on societal issues.

BY JULIA ROSEN

Megan Evans got a crash course in science policy in 2011. As a research assistant at the University of Queensland in Brisbane, she joined a project helping the Australian government to develop a tool to compensate for the environmental effects of commercial land development and other activities. If a protected species might be harmed, for example, the 'biodiversity offset' tool would help the government to determine how much extra habitat to set aside. Evans loved the project's applied nature.

Many early-career researchers are drawn

to the intersection of science and policy, says Evans, now an honorary research fellow at the Centre for Policy Futures at the University of Queensland. But it can be hard to know where to start, she says. And there can be career penalties for junior scientists. Policy-based work can be time-consuming and hard to fund, and helping to shape a law or management plan might not look as good on a tenure application as do high-profile publications. All scientists must also cope with the political realities of helping to translate scientific evidence — replete with uncertainties — into clear-cut laws and regulations. Because of this, many say, science can underpin good policy, but rarely defines it.

Even so, engaging in policy has never been more important, says Tateo Arimoto, a science-policy expert at the National Graduate Institute for Policy Studies in Tokyo. Society and the world are changing rapidly, he says, and policymakers need scientific evidence to guide decisions on issues from climate change to artificial intelligence. "The mission of modern science is not only creating new knowledge," he says, but "using scientific knowledge to address social issues".

Researchers can take proactive measures to increase the policy impact of their work. They should establish strong relationships with elected officials or government staff members, and learn to provide clear and concise summaries of existing scientific evidence to help policymakers to understand the options. Scientists and policymakers can also collaborate on projects aimed at real-world questions. The important thing is to be humble and open, Evans says. "If you want to engage with policy, you need to go cap in hand, and say, 'How can I help?'"

CONNECT AND OBSERVE

The first step, Evans says, is to connect with policymakers. In a paper this July designed to help other early-career scientists to navigate the policy landscape¹, Evans and Chris Cvitanovic, a researcher at the University of Tasmania's Centre for Marine Socioecology in Hobart, suggest that scientists first observe how policymaking works for their issue of interest. Approaches such as reading the news and setting up Google alerts for relevant keywords are helpful, they say.

Then, scientists can determine who in the policy world might be interested in particular aspects of their work and why, and how those people interact with one another. Lawmakers, officials in a national government's executive branch and their aides could be one audience, as could staff members at government agencies who implement those policies. Evans recommends sketching a map of potential contacts that researchers can refine over time.

Senior scientists with existing policy contacts can help early-career researchers to make connections. Scientists can also introduce themselves and their work to the legislators who represent their home districts. "It can be as simple as getting out of the office and going to talk to people face to face," says David Rose, an environmental geographer at the University of East Anglia in Norwich, UK, who studies science and policy. He also advises scientists to contact groups of lawmakers who are interested in the issues they study. For instance, ►

► members of the US Congress have created caucuses, or alliances, to advance neuroscience and planetary science. The United Kingdom has all-party parliamentary groups on such topics as cancer and wildlife conservation, and, in Australia, parliament has 'friendship groups' focused on science and medicines.

Rose also recommends setting up meetings with government employees who provide science advice to lawmakers, such as members of the European Union's Parliamentary Research Service, or government science advisers. Peter Gluckman, who was chief science adviser to the prime minister of New Zealand until June 2018, says that for maximum impact, written letters highlighting an issue or providing science advice should come from a professional society, institute or national academy. Still, blogging and using social media can increase visibility for scientists and the issues they want to emphasize, Evans says, and Twitter can help in connecting with key policymakers.

Researchers might also forge fruitful relationships with employees of the government agencies and departments that work to enact existing legislation. For example, California laws require the state to reduce its greenhouse-gas emissions by 40% below 1990 levels by 2030, partly by storing more carbon in soils. So Katharine Mach, a climate-assessment scientist at Stanford University in Stanford, California, has been helping the state's agriculture, forestry and other agencies to evaluate the benefits of land-management practices such as adding compost or charcoal to soils.

Mach and her colleagues joined the effort at the invitation of the S. D. Bechtel, Jr. Foundation and the David and Lucile Packard Foundation, both in California, which sought the researchers' expertise in policy-relevant climate science. But Mach says that scientists at any career stage can help to shape government programmes. One effective way, she says, is to submit letters and evaluations when officials solicit public feedback on proposed regulations or plans of action. "Those are incredibly important and also kind of fun," she says. "You are thinking in real time about a good approach." She signs up to government e-mail lists to stay apprised of upcoming workshops and requests for input. (Alternatively, Evans says, researchers can make connections by offering to give a talk at an agency or in a department's regular seminar series.)

MEET AND GREET

Toni Lyn Morelli, an ecologist at the US Geological Survey in Amherst, Massachusetts, recommends attending a variety of conferences. She wanted to connect with state wildlife officials about her work on the future of streams in which cold-water fish live. She decided against organizing a session at the annual meeting of the Ecological Society of America because she knew that few managers would attend. So she went to a conference hosted by the Northeast Association of Fish and Wildlife Agencies, where she

reserved a room and invited managers to stop and talk — and eat pizza. "We got great people."

When scientists get involved in policy, they should be careful not to advocate for specific solutions, warns Gluckman. Instead, he says, quoting from a book by political scientist and public-policy expert Roger Pielke Jr, a scientist should be an 'honest broker', helping policymakers to understand possible policy options and their consequences.

This was Craig Downs's approach when he helped Hawaiian legislators to draft a bill to ban sun creams containing chemicals that research from Downs and others has shown to be harmful to coral reefs². Downs, an ecotoxicologist and director of the non-profit Haereticus Environmental Laboratory in Clifford, Virginia, explained to lawmakers the chemicals' impacts and the implications of policy options, such as imposing a temporary or a permanent ban, but didn't advocate for one in particular. He knew that legislators had to balance many factors, including how the ban might affect sun-cream manufacturers. (Facing strong public pressure, the lawmakers passed a permanent ban in May. It was approved last month.)

In any interaction, Rose says, it's important to use clear, accessible language and, if possible, to tell a compelling story about the science. Most of all, scientists should understand that policymakers rarely want to hear about the results of a researcher's latest peer-reviewed study. When Rose polled members of the UK Parliament, he found that most wanted a succinct overview of the current body of knowledge on an issue³. Arimoto says that researchers should try to bring in as many threads as possible that might be relevant to policy. "Individual scientists need not only the capability of analysis, but also to synthesize," he says.

Downs suggests honing a three-minute 'elevator pitch' for in-person meetings with

lawmakers. Gluckman advises scientists to prepare written materials as policy briefs, leading with key points, offering relevant caveats and then laying out possible options. (Johns Hopkins University in Baltimore, Maryland, offers an online guide (go.nature.com/2puyq35); researchers can also contribute to scientific reviews targeted at policymakers, such as those published by the Oxford Martin School, UK, and the Campbell Collaboration in Oslo.)

Scientists can seek in-depth training on how to interact with policymakers. Gluckman chairs the International Network for Government Science Advice, which hosts conferences and workshops that bring together scientists and policymakers worldwide. Many universities and professional organizations, including the American Institute for Biological Sciences in McLean, Virginia, offer 'boot camps' for researchers.

Gluckman also recommends that scientists take a sabbatical in the policymaking sphere. For instance, one can apply to be a Science and Technology Policy Fellow with the American Association for the Advancement of Science, or to be a research fellow at the European Commission's Joint Research Centre. Scientists can also take a temporary appointment at a government science agency, the United Nations, the World Health Organization or the Organisation for Economic Co-operation and Development (OECD), among others. Those who have policy experience, Gluckman says, learn how to operate in both worlds.

SLOW BURN

Scientists who engage in policy should not expect immediate results. The diffusion of science into policy is often incremental, says Matthew MacLeod, an environmental chemist at Stockholm University. His research group is designing a new version of the test that the OECD recommends countries use to assess bioaccumulation of a substance when deciding how to regulate it. His version takes less than half the time of the standard test and requires

"Science can underpin good policy, but rarely defines it."



Japanese science-policy expert Tateo Arimoto uses scientific knowledge to address social issues.

about one-third of the fish, which serve as the test subjects. But he anticipates that it will be ten years before it's adopted.

Often, a catalysing event piques policymakers' appetite for scientific evidence. That's why scientists should make a long-term investment in policy work, Evans says, and be ready to act when the opportunity arises. For instance, she recalls, the Australian government decided to implement the biodiversity-offsets project when a new minister took office, and drew on well-established research. "We ended up being able to use that science really quickly," Evans adds that researchers should pay attention to changes in administrations in their own and other jurisdictions that might increase the receptiveness of policymakers to scientific evidence.

There can be cases, however, when the evidence isn't yet strong enough to spur action, says Ian Boyd, chief scientific adviser at the UK Department for Environment, Food and Rural Affairs. For example, he says, research over the past decade on whether neonicotinoid pesticides harm bees hasn't yielded clear answers about population-level effects. In an opinion article earlier this year, Boyd explained he had become convinced that the chemicals were being used more widely than was recognized and offered growers only a marginal benefit⁴. However, he lamented the lack of rigorous studies quantifying the actual danger they posed to pollinators. The United Kingdom ultimately backed the EU's decision to ban the chemicals.

To make sure science influences policy, it's best to collaborate with policymakers from the start, says Mach. "Scientists doing science in isolation won't know what questions are most relevant, and also won't really influence decisions," she says. Collaboration requires reaching out to policymakers and agency staff long before research begins, listening closely to their questions and needs, and shaping studies around those. After that, she says, scientists must maintain regular contact, share preliminary results and be ready to change the focus of a research project in response to feedback.

It's challenging, but Mach and others find working at the interface of science and policy extremely rewarding. After all, like many researchers, Mach went into science eager to tackle issues that matter. "There's something that's really motivating about doing science that is attuned to the bigger picture," she says. ■

Julia Rosen is a freelance writer in Portland, Oregon.

1. Evans, M. C. & Cvitanovic, C. *Palgrave Commun.* **4**, 88 (2018).
2. Downs, C. A. et al. *Arch. Environ. Contam. Toxicol.* **70**, 265–288 (2016).
3. Rose, D. C. *Br. Ecol. Soc. Bull.* **48** (4), 34–35 (2017).
4. Boyd, I. L. *Nature Ecol. Evol.* **2**, 920–921 (2018).

COLUMN

Stronger together

Lab heads should foster collaborative research, say
Katherine D. Kinzler and Kristin Shutts.

Two of our PhD students were in a bind. They had collaborated on a research project that merged their interests and, as counselled by other faculty members, had decided early in the research process on authorship order. But by the end of the partnership, the designated second author felt that she had contributed more time and expertise to the project, and wanted to switch the authorship order. The would-be first author disagreed, pointing to their earlier arrangement. Disappointment, or worse, seemed the probable outcome.

This scenario might feel familiar to many principal investigators (PIs). At best, considering contribution and authorship order can be stressful for students and postdocs who collaborate; at worst, these issues can prevent alliances from developing at all. Yet, in our experience, as student collaborators ourselves and then as PIs, some of the best science — and the impetus for growth in junior researchers' careers — comes from collaborative efforts between graduate students and/or postdocs. As PIs, we work to set the tone for joint science to flourish in our labs.

We began our own collaboration as PhD students in the same lab at Harvard University in Cambridge, Massachusetts. Working together has produced positive outcomes for both of us — from developing more-advanced records for the job market (then) to receiving a multi-year federal grant from the US National Institutes of Health that we jointly administer (now). Most importantly, we've come to believe that the ideas we generate as a two-person team are better than what either of us would produce alone, and that the scientific process is more fun to conduct together.

Consequently, we were surprised to encounter push-back when we suggested in our own labs that students consider working together. So, we developed a model to foster collaboration.

Eliminate a 'zero-sum' mindset. Collaboration can help to direct students to 'growing the pie' — creating more resources together that they can ultimately share. As graduate students, we developed a shared research programme that generated multiple studies and articles, so determining authorship was never stressful for us. We encouraged the students in the anecdote above to think about generating a pipeline of collaborative projects. By treating the project as the first step in an important, long-term programme, neither student felt as worried about the final authorship decision.



Establish parameters. Recently, a new student in one of our labs wanted to collaborate with a postdoc, yet devoted significant attention to dissecting her role in the project and how much time she (compared with the postdoc) was spending on it. All this worry risked stagnating the science and ending the collaboration. We explained the benefits of this type of partnership, and pointed to how our own successes, as well as those of previous students, have been bolstered by sharing credit with other scientists.

Encourage students to make authorship decisions after they collect data. In our experience, determining authorship later in the process puts the science (rather than the publication process) front and centre, and helps students to think of growing the total amount of research, rather than angsty over whether they plan to contribute 49% or 51% of any given project.

Of course, we recognize that collaboration might not work for all student pairs. Collaborative relationships, in our experience, are most likely to flourish when junior researchers lead them. PIs should help students and postdocs understand the value and process of collaborative work. But junior scientists should initiate specific collaborative projects and decide together how to carry out the research.

"Let's put the science ahead of ourselves," agreed our two students deciding on authorship order. One was first author on the initial paper — which sparked a new research programme — and the other was first author on a subsequent publication. Collaboration benefits both the students and the science. ■

Katherine D. Kinzler is a psychologist at Cornell University in Ithaca, New York.
Kristin Shutts is a psychologist at the University of Wisconsin–Madison.

UNREAL

The possibilities are endless.

BY JUDY HELFRICH

“Every time it’s them same damn nuts.” Mabel narrowed her eyes at the nut bowl on the coffee table. “D’you suppose it’s the same nuts what’s been out the last ten years?”

“One way to find out,” said Bertha. “Try one.”

The two sisters regarded the nuts.

“Nah,” said Mabel. “I ain’t that brave.”

“I’ve a hankering,” Bertha bit into a filbert, then raced into the kitchen, retching.

Iris emerged through the kitchen door. “What’s with her?”

“She had one a your nuts,” said Mabel. “How old would them things be?”

“Don’t matter how old. They ain’t real.”

“What are you on about?”

“Them scientists on the radio,” said Iris. “They’re in a tizzy about some discovery, saying none a this is real. That we’re playing out every little thing that never happened instead.”

“That’s a load a hooley,”

Bertha weaved in from the kitchen and toppled into an armchair, clutching her stomach.

“That’s how I felt last time I ate here,” said Mabel.

“Now I know this is unreal,” Iris said. “Cause I’d never have you two over in real life.” She huffed back into the kitchen.

A head popped through the living-room wall, to the left of the television. It leered at Bertha and Mabel with a sloppy grin and cloudy blue eyes.

Bertha surveyed the head with disgust. “Not only are you poisoned when you eat here, you got to put up with heads.”

Mabel tut-tutted. “You get them heads if you don’t keep a proper house.”

Bertha nodded. “You got to get after them heads; nip them in the bud.”

Another head broke through, just below the first.

“She’s lucky she don’t got blonds,” said Mabel. “Them redheads ain’t so bad.”

A blond head burst through the wall.

“Speak a the devil,” said Mabel.

Iris bustled into the living room. “Now them scientists are saying we’re in an unreality loop and things are gonna get weirder and weirder.”

“Iris,” said Bertha. “Don’t you think we’d notice if things was getting weird?”

“Heads!” Iris squawked. She rushed for a broom and beat the heads about the head.

The redheads screwed up their faces and withdrew. The blond took a bit more beating before it threw her a baleful look and popped back into the wall.

“You know where them heads come from,” said Mabel.

Iris eyed her.

“Spoilt food. Attracts ‘em.”

“Like them 20-year-old nuts,” said Bertha.



“It’s a wonder them heads ain’t thick like them roaches you had last year,” said Mabel.

“I never had roaches,” Iris spat.

“Member, Bertha? Iris left them dates out and you had half a one before it crawled away.”

“Yeah,” said Bertha. “Best thing I ever ate here.”

The two sisters cackled.

A brunette head popped through the coffee table and began feasting on the nuts.

“What’d I tell you,” said Mabel.

Iris issued a smack to the head. It growled at her. She whacked it a good one and it withdrew.

“You can’t go about thumpin’ them heads,” said Bertha. “You need proper head poison.”

“And for every head you see, there’s 50 more behind your walls,” said Mabel.

“You can start by tossin’ them nuts,” said Bertha.

“Them nuts is tradition,” said Iris.

“So’s the ptomaine you give us, but you don’t see us pining for it.”

A public-warning siren blared outside.

“Now what?” said Mabel. She clicked on the TV.

➔ **NATURE.COM**

Follow Futures:

🐦 @NatureFutures

📘 go.nature.com/mtoodm

“... stranger and stranger. For everything that happens,” said a scientist,

“almost infinite things don’t happen. Except they do. Everything that doesn’t happen in reality happens in unreality. That’s here. That’s us. We’re unused quantum superpositions. Uncollapsed wavefunctions. Unreal.” Bats blasted out of his ears. The scientist stared at them, wild eyed. “Is that normal? Because I can’t tell.” He drummed his fingers against a wandering dodo. “We can expect events to become more and more implausible until we exhaust every single possibility and the unreality loop collapses.”

“Pfft.” Mabel clicked off the television. “Our tax dollars at work.”

“You can’t keep 30-year-old nuts about,” said Bertha. “People want to sample them things, like I did.”

“Least it didn’t have legs,” said Mabel.

“You sure?” asked Bertha. “Them nuts is so old I thought they evolved some.”

Iris stormed back into the kitchen.

“You know,” said Mabel, observing a new head. “That one kinda looks like you.”

Another head popped through the wall.

“And that one looks like you,” said Bertha.

The Mabel-head bit the Bertha-head. Bertha grabbed Iris’s broom and walloped it. It gnashed at the broom and got a hold with its teeth. Bertha yanked the broom free, falling backwards and bashing her head against the coffee table.

The Bertha-head’s shoulders emerged from the wall.

Bertha twitched. She faded. She disappeared. New Bertha climbed out of the wall. She leered and shuffled towards Mabel.

Mabel pelted it with nuts. A filbert bounced off its eye, but it seemed not to notice. It grabbed Mabel by the throat and throttled her until she faded away.

New Mabel emerged from the wall.

Iris banged in from the kitchen.

The two of them faced Iris. They leered at her.

Iris leered back.

Then they sat around the coffee table and played pinocle.

Iris won. She ate a nut to celebrate.

“Every time,” said Mabel. “It’s them same damn nuts.” ■

Judy Helfrich is a writer and visual artist who hails from the Canadian prairie, where long stretches of nothing persist in at least four dimensions. More at helfrich.ca.

ILLUSTRATION BY JACEY